Universität
Augsburg
University

Social Augmentation Using Behavioural Feedback Loops

Ionuț Damian

Dissertation for the academic degree of

Doctor rerum naturalium
(Dr. rer. nat.)

submitted to the Department of Computer Science

Lab for Human Centered Multimedia
University of Augsburg

2017

Universität
Augsburg
University

Date of examination:	18.07.2017
Reviewers:	Prof. Dr. Elisabeth André
	Prof. Dr. Bernhard Bauer
	Prof. Dr. Johannes Schöning

Abstract

Recent technological advancements have enabled the miniaturization of electronics to a degree that powerful mobile computers and versatile sensor arrays can be integrated in phones, watches or even glasses. This dissertation looks at how such devices, in particular smartphones, smart watches and smart glasses, can be used to help users participating in social interactions improve the quality of their behaviour, and thus better the outcome of the interaction. To this end, the concept of social augmentation is introduced. Social augmentation makes use of behavioural feedback loops to analyse the behaviour of the user in realtime and, based on the outcome of this analysis, provide live feedback to the user on how to improve their behaviour. The concept is positioned as an evolution of classical social skills training methods, capable of augmenting the social skills of the users in realtime during actual social interactions. It is targeted at users who wish to perform better during critical social interactions, such as job interview or speaking in public. However, social augmentation can also help persons who suffer from various disabilities better regulate their behaviour to avoid misunderstandings and generally increase their functional independence.

The main challenge of this approach lies in designing the augmentation to function alongside social interactions without disrupting them. To address this, the thesis introduces a conceptual framework for social augmentation, which has been informed by empirical and theoretical findings from the fields of cognitive psychology, human-computer interaction and digital signal processing. With the help of three user studies spanning two scenarios (public speaking and group discussions), the social augmentation concept has been tested for effectiveness in improving social behaviour without disrupting the social interaction.

To foster further research in this area, the SSJ open source software framework for social augmentation has been implemented. Through a combination of state-of-the-art behaviour analysis and live feedback techniques, SSJ supports the behavioural feedback loop in its entirety. First, it allows the processing and classification of social signals extracted from various sensors locally on mobile devices and in realtime. Using the results of the behaviour analysis, SSJ can deliver unobtrusive, multimodal live feedback using various output devices including head-mounted displays, vibro-actuators and headphones. Thus, SSJ enables the creation of powerful and versatile social augmentation systems. Moreover, thanks to a modular and flexible design, the augmentation is not restricted to any particular scenario, but can be targeted at manifold social situations.

Keywords: Social skills training, mobile social signal processing, multimodal live feedback

Zusammenfassung

Der rasante technische Fortschritt aus dem letzten Jahrzehnt ermöglichte die Entwicklung von immer kleiner werdenden elektronischen Geräten. Mittlerweile können leistungsstarke Prozessoren und präzise Sensoren in Telefonen, Uhren und sogar Brillen verbaut werden. Im Rahmen dieser Dissertation wird untersucht, inwiefern das Nutzerverhalten in einer sozialen Interaktion unter Ausnutzung der Sensorik und Aktuatorik miniaturisierter mobiler Geräte verbessert werden kann. Dazu wird das Konzept der "sozialen Augmentierung" eingeführt, das eine Verhaltensrückkopplungsschleife (Engl. behavioural feedback loop) einsetzt, um den Nutzer bei der Verbesserung seines sozialen Verhaltens kontinuierlich zu unterstützen. Hierbei wird zunächst das Verhalten des Nutzers in Echtzeit analysiert und mit Hilfe von digitalen Signalverarbeitungsmethoden ausgewertet. Basierend auf dieser Analyse gibt das System dem Nutzer individualisiertes Feedback, das durch Anwendung zu einer Verbesserung des Verhaltens führen soll. Durch eine stetige Wiederholung dieses Vorgangs soll das Verhalten so angepasst werden, dass eine deutliche Verbesserung der sozialen Fähigkeiten, wie Gestik, Mimik, paralinguistische Merkmale u.a. eintritt. Positioniert wird dieses Konzept als eine Evolution der klassischen Trainingsmethoden für "Social Skills." Im Gegensatz zu diesen, wird das Verhalten des Nutzers in Echtzeit, d.h. während der eigentlichen Interaktion verbessert ("augmentiert"). Als Einsatzgebiet von sozialen Augmentierungssystemen werden vor allem anspruchsvolle soziale Situationen anvisiert, wie zum Beispiel, Vorstellungsgespräche oder öffentliches Reden. Dabei soll der Nutzer unterstützt werden, ein besseres Ergebnis bei solchen Interaktionen zu erzielen. Auch für Personen mit Beeinträchtigungen könnten Augmentierungsmethoden eingesetzt werden, um ihnen zu helfen, sich während alltäglicher sozialen Situationen besser auszudrücken und somit Missverständnisse zu vermeiden.

Die größte Herausforderung ist es, die Augmentierung so zu gestalten, dass sie während einer echten sozialen Interaktion effektiv ist, ohne dabei den Nutzer, die Interaktion oder die Interaktionspartner zu stören. Zur Lösung dieser Herausforderung wurde aus Erkenntnissen der Kognitionswissenschaften sowie Theorien aus der Mensch-Technik-Interaktion ein konzeptionelles Framework für die Augmentierung von soziale Interaktionen erstellt. Dieses wurde mit Hilfe von drei Nutzerstudien evaluiert und auf Effektivität geprüft. Dafür wurden Systeme zur Augmentierung von öffentlichem Reden und Gruppendiskussionen implementiert und unter realen Bedingungen getestet.

Um die weitere Forschung im Bereich der sozialen Augmentierung zu fördern, wurde ein frei verfügbares technisches Framework namens SSJ entwickelt. Durch die gezielte Kombination von Signalverarbeitungsmethoden und Live-Feedback-Techniken, ermöglicht SSJ die Erstellung von Augmentierungssystemen für unterschiedliche sozialen Situationen. Genauer gesagt erlaubt SSJ die Aufzeichnung, Verarbeitung und Klassifizierung von

Sensordaten auf mobilen Geräten und in Echtzeit. Zusätzlich können mit Hilfe von SSJ komplexe Feedback-Strategien definiert werden, welche die automatische Generierung von multimodalem Feedback in Reaktion auf Nutzerverhalten ermöglichen. Für Datenerfassung und Feedbackübermittlung, unterstützt das Framework eine große Anzahl an Sensoren und Ausgabegeräten.

Schlagwörter: Training von Social Skills, mobile Verarbeitung sozialer Signale, multimodales Echtzeit-Feedback

Acknowledgements

First of all I would like to express my deepest gratitude towards my supervisor, Prof. Dr. Elisabeth André, for her support and guidance over the entire PhD program. Her constant advice not only steered my research efforts but also shaped this dissertation into the coherent piece of work now laying in front of you. Furthermore, I thank Prof. Dr. Johannes Schöning for his counsel and constructive feedback on numerous occasions. I am also grateful to Prof. Dr. Bernhard Bauer for agreeing to review this dissertation.

Furthermore, my thanks also go to my colleagues, who made working at the lab such a pleasant experience, and to the various students which contributed to my research. I am especially thankful to Birgit Lugrin for guiding me during my undergrad years, Tobias Baur for working on TARDIS with me and co-authoring the vast majority of my papers, and to those who helped with the development of the SSJ framework: Michael Dietz, Frank Gaibler and Simon Flutura.

Finally, I am deeply grateful to my parents for putting me on this journey and supporting me every step of the way, and to my brother, Bogdan, for always being there for me. Last but not least, I would like to thank Ana for her endless love and for making me believe in myself.

Contents

List of Figures

List of Tables

1. Introduction

Social behaviour lies at the very core of being human. We engage in social interactions multiple times every day. For example, we speak with friends, buy products from a salesman or hold conversations with colleagues. Yet, some types of social interactions (e.g. speaking in public, being interviewed, participating in a negotiation) often seem to be governed by a special set of rules, many of us struggle with. However, specifically these kinds of interactions are the ones where the outcomes are the most crucial. For example, a job interview can decide our employment status, or a presentation in school may have a large impact on the final grade. The problem is amplified by the fact that in such critical situations, stress can make our bodies go into "auto pilot" mode, rendering our cognitive minds oblivious to our body's use of gestures, postures or even speech. I found this effect particularity obvious during one of my recent conference presentations when, about 10 minutes in, I noticed that the session chair was vigorously gesticulating and mouthing "look up!" It was at this moment that I realized that for the past 10 minutes I have been staring at the floor instead of making eye contact with the audience. The same effect can also be seen in the case of habitual behaviour (also called mannerisms) which are often performed without us being consciously aware of it. For instance, my default body posture has always been "droopy", i.e. slightly bent back, lowered shoulders, gaze directed downwards. Although I have been trying to address this for the past several years, my corrections are short-lived: After a certain amount of time I always go back to my usual droopy stature. Besides having some obvious aesthetic disadvantages, I noticed that my body posture sometimes causes me to slide in a submissive role during social interactions. In such situations, I often wish someone, or something, could stand next to me, to make me more aware of my own behaviour and how it might be perceived by others, to give me feedback on what I am doing wrong and tips on how to do better.

Imagine a computer system able to monitor our behaviour during social interactions and give us subtle feedback on how to improve. For example, when speaking in public, such a system could help us maintain eye contact with the audience or control our speaking rate by informing us if we speak too slow and boring, or too fast and unintelligible. The same system

could also help us during job interviews make a better impression by correcting our body posture and use of gestures. Such a system would effectively "augment" the user's social skills.

This thesis introduces the novel concept of *social augmentation*. In social augmentation, behavioural feedback loops continuously analyse the behaviour of the user and, based on this analysis, live feedback is provided to the user with the aim of improving their social behaviour. For this, physical artefacts, such as miniaturized sensors or light-weight displays, are used to deliver feedback to the user while they are engaged in real social interactions. Social augmentation can be seen as an extension of Engelbart's framework for augmenting human intellect [Engelbart, 1962] and the personal augmentation concepts of Xia and Maes [2013]. It is based on the consideration that social behaviour is at the core of human intellect. Earlier studies by Scherl and Haley [2000] have shown that the display of conversational aids during social interactions may improve communication. Furthermore, studies by Ofek et al. [2013] seem to indicate that secondary information can be consumed by users during a conversation without the interlocutors noticing it. Despite encouraging findings, there are also qualified concerns that social communication might be impaired by the additional information the user needs to process. For example, there is a risk that the augmentation system could lead to reduced gaze contact because the user needs to split their attention between the feedback and the social interlocutor. Studies by McAtamney and Parker [2006] point out some issues which arise when users need to divide attention between a social interaction and a secondary task: While wearing a non-activated head-mounted display does not seem to influence how others perceive the user, an activated display may have a negative impact on the social interaction. Thus, to allow the augmentation to function alongside live social interactions, this thesis proposes the use state-of-the-art mobile social signal processing techniques for "in the wild" behaviour analysis, and intelligent feedback strategies for delivering subtle yet compelling feedback to the user.

1.1 Research Objectives

Several research objectives are targeted by this dissertation. These range from conceptual to technical and empirical.

- The first research objective of this thesis is to provide a conceptual framework for social augmentation. From the few existing examples of social augmentation-like systems [Boyd et al., 2016; McNaney et al., 2015; Tanveer et al., 2015], which have been developed concurrently to this dissertation, one can notice a general confusion about how to approach the topic, as highlighted by conflicting terminology and unfounded design decisions for feedback strategies. A conceptual framework would support the design and development of social augmentation systems by introducing clear guidelines and requirements. This would also foster future research in this domain by providing a common basis for discussion and allowing for comparison between different social augmentation approaches.
- A social augmentation system must be able to work during actual social interactions. Thus, the second research objective is to introduce a behaviour analysis solution which is able to perform online processing and classification of user behaviour in mobile and uncontrolled environments. However, while the field of social signal processing has seen a lot of progress in the recent years, offline behaviour analysis and classification

is still the method of choice for most researchers. Wagner's survey [Wagner, 2016] found that only 10% of all social signal processing papers listed on the SSPNet portal[1] attempt realtime analysis of social signals. Progress in this area would not only benefit the social augmentation concept, but also support the development of mobile realtime applications capable of reacting and adapting to the user's behaviour "in the wild."

- The third research objective is to find a solution for delivering live feedback to the user without interfering with the main task, i.e. the interaction itself. In classical social skills training approaches, the training usually happens offline during dedicated sessions. This is for good reason: It allows the user to fully concentrate on the learning and knowledge acquisition. What social augmentation attempts is to have the training happen simultaneously to the participation in a social interaction, thus demoting it from primary to secondary task. A system capable of delivering feedback under these circumstances needs to be mindful of the fragile nature of human attention and able to automatically adapt to the user, the moment and the situation.
- Finally, the introduced concepts need to be evaluated to determine their effectiveness in improving user behaviour. Yet, due to the focus of social augmentation on working alongside actual social interactions, it is paramount that evaluations also happen under realistic conditions. Isolated, controlled studies are likely to be insufficient for this task. The overall aim of such evaluations is to demonstrate whether the social augmentation is capable of altering user behaviour, and whether it is able to do so without disrupting the user, the interlocutors and the flow of the social interaction. Besides answering questions regarding the performance of the social augmentation concept, empirical findings of this kind would also advance our knowledge of how technological system integrate into day-to-day lives.

1.2 Outline

The dissertation is split into four parts and ten chapters. This structure is illustrated in Figure 1.1.

Part one first aims to familiarize the reader with the main ideas and theories related to social interactions, and gives a general understanding of how human attention works and how it can be distributed among different tasks (Chapter 2). Chapter 3 then provides an overview of the field of social skills training. To exemplify current state-of-the-art in terms of computer-enhanced training, one concrete implementation of such a system is presented. Using this example as a starting point, the social augmentation concept is positioned as an evolution of computer-enhanced training.

The second part introduces the social augmentation conceptual framework. First, an overview of social augmentation, its requirements as well as its components is provided (Chapter 4). Moreover, the behavioural feedback loop is introduced as the driving force behind the social augmentation concept. Then, Chapters 5 and 6 present the two main components of a behavioural feedback loop: mobile social signal processing-driven behaviour analysis and multimodal live feedback generation. Using extensive literature reviews, the design spaces for both components are explored and discussed.

In the third part, concrete implementations of the social augmentation concept are provided. In Chapter 7, a novel software framework for designing, creating and executing behavioural

[1] `http://sspnet.eu`

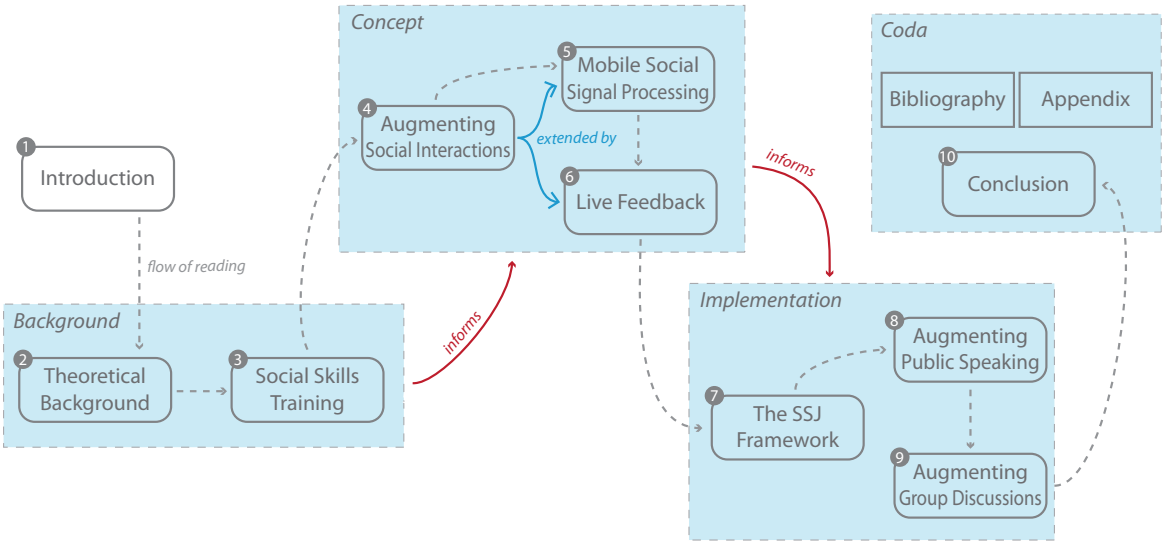

Figure 1.1: Thesis structure.

feedback loops with the help of modern wearable devices is introduced. The framework realizes both pillars of a behavioural feedback loop: behaviour analysis using mobile social signal processing techniques and live feedback generation. Then, two social augmentation applications are described. First, Chapter 8 presents a system for augmenting speaking in public with the help of a Google Glass[2] head-mounted display. Second, a face-to-face discussion augmentation system is described in Chapter 9. Both systems are evaluated with the help of user studies, from which conclusions are derived for the merit of the overall social augmentation concept.

Finally, in Part four, the thesis is concluded, the contributions of the work are discussed and an outlook for future research is provided. This part also contains the bibliography as well as the appendices, which include supplemental material the interested reader might find useful. An overview of all personal publications and how they contributed to the dissertation is provided in Appendix H.

[2] https://en.wikipedia.org/wiki/Google_Glass

I Background

2. Theoretical Background

The aim of the present dissertation is to introduce an approach for social augmentation in an effort to improve the quality of the user's behaviour during social interactions. This chapter provides the reader with background information on the supporting pillars of this concept. More specifically, it starts with a discussion on what social behaviour is, what its role is during social interactions and what happens when social behaviour is used inappropriately.

The second part of the chapter presents various theories of cognitive psychology regarding human attention. These theories are important for understanding how the social augmentation task may impact the performance of the social interaction itself. Thus, a close look will be given to how attention can be divided across multiple tasks, and what happens when the user is overwhelmed and the tasks get interrupted.

2.1 Social Interaction

When interacting with each other, humans use multiple types of communication to transfer information. The most obvious one is the verbal communication. Verbal communication conveys meaning in a very direct way. The sender consciously encodes the information using the rules of a preferred language into sequences of words and then vocalizes these using the human vocal apparatus. On the receiving side, the messages are perceived using the auditory system and then decoded using (hopefully) the same rules the speaker used.

Nonverbal communication is the process through which humans exchange information without the use of speech. It includes the use of body language (kinesics), distance (proxemics), touch (haptics) and gaze (oculesics). A typically overlooked form of nonverbal communication is paralinguistic behaviour, i.e. how one speaks (e.g. intensity, speech rate, voice modulation).

While most of the time we are only aware of our verbal behaviour, both forms of communication – verbal and nonverbal – are critical to our day to day interactions. Even when we simply tell a person on the street the time, we transmit dozens of nonverbal behaviours

in parallel to the verbal message. We might smile or frown, make eye contact or look away, change our posture, raise our shoulders, speak fast or slow, high or low; and so on. We often transmit all of these nonverbal messages unconsciously to our interlocutors, and they also often perceive, interpret and respond to them just as unconsciously. If we smile, the other persons might smile back, they will respond to our eye contact by either looking at us or away, and they might adapt their posture to match ours.

However, inappropriate use of nonverbal cues can lead to unwanted effects. Our social lives are governed by a set of unwritten rules, called social etiquette. These rules change from context to context, from situation to situation. For example, it is accepted to gaze at other peoples' faces in public but only briefly. Lengthy directed gaze (starring) is considered rude and aggressive behaviour. In a job interview, it is expected of the interviewee to behave in a calm and composed fashion, agitated gesturing may negatively impact the interviewee's chance for employment [Hollandsworth et al., 1979]. Similarly, while giving a public speech, inappropriate nonverbal behaviour (e.g. excessive gesturing, freezing up, speaking too fast) caused by stress or nervousness is considered bad practice as it impacts the way the audience perceives the speech.

In the following pages a closer look at nonverbal behaviour will be provided. For this, the different types of nonverbal behaviour are presented and their role during social interactions discussed.

2.1.1 Kinesics

Kinesics refers to nonverbal communication using parts of the human body, or the body as a whole. It comprises gestures, postures and facial expressions.

Ekman and Friesen [1969] categorized gestures into 4 categories: emblems, illustrators, regulators and adaptors. Each category is used with a specific communication function. Emblems are culture-specific hand gestures that hold a special meaning (e.g. thumbs up). Illustrators are gestures used to complement or illustrate the verbal part of the communication. Regulators are culture specific gestures used to regulate the flow of the communication. Adaptors are gestures or habits used as reactions to internal or external stimuli (e.g. covering your mouth, scratching your nose). McNeill [1992] refines this categorization by further splitting illustrators and regulators in: iconic, metaphoric, deictic and beat gestures. Iconic gestures are gestures that have a close relationship with the utterance and are used to visually complement it (e.g. using two fingers to visualize a walking movement). The metaphoric gestures visualize abstract concepts with the help of metaphors (e.g. showing a spherical shape when talking about one's accomplishments). Deictic gestures define references in space, for instance when someone points towards the door and asks another person to leave. The references can also be abstract such as when someone points to the floor and says "you're going down." Finally, beat gestures are rhythmical movements that emphasize certain parts of an utterance.

An adequate use of gesture is not always easy since many gestures are culture-specific. For example the gesture where the tip of the thumb touches the tip of the index finger symbolizes "ok" in the US and Europe but is an insult in Brazil meaning "You are all a bunch of arseholes" [Pease and Pease, 2008]. The most famous incident was Richard Nixon's visit to Brazil in the 1950s. When getting out of the air plane, he looked at the cameras and performed the "ok" gesture to the shock (and amusement) of the Brazilian population.

The second big part of kinesics is postures. A posture is defined as a specific configu-

ration of the limbs. Postures have been found to communicate openness, confidence and engagement [Pease, 1988]. Generally, open postures are considered warmer and indicate one's willingness to cooperate, whereas closed postures do the opposite. Certain postures are linked to specific messages. For example, placing both hands behind the head signals confidence and dominance. However, Mehrabian argues that the effect of postures depends on the social status of the interlocutors [Mehrabian, 1969]. The fact that postures are often used unconsciously is particularly apparent when looking at posture mimicry, also called *the chameleon effect*. It involves the involuntary copying of another person's behaviour, including their posture, during a social interaction. Chartrand and Bargh [1999] found that "mimicry facilitates the smoothness of interactions and increases liking between interaction partners."

Lastly, facial expressions represent patterns of movements in the human face and are closely linked to human communication of emotion. While most of us can only name several emotions or facial expressions, Ekman found that there are over 10000 different facial configurations [Ekman, 2003]. Here, it is important to note that while humans use facial expressions to communicate emotions, these do not necessarily reflect the internal feelings. During social interactions, facial expressions are mostly used voluntarily according to one's social goals [Fridlund, 2014]. For example, we smile when we want to signal happiness and frown when we want to signal sadness. Nevertheless, facial expressions can also be involuntary. Internal mental and emotional states can manifest themselves as facial expressions without us being aware of it. Such involuntary facial expressions are referred to as microexpressions [Ekman, 2009]. Besides display of emotions, facial expressions can also act as a secondary communication channel. Smiling can signal respect for the conversation-partner or interest in another person. Moreover, facial expressions can also act as emblems which hold specific meaning in certain cultures [Ekman and Friesen, 1975]. For example, we wrinkle our nose when we want to signal disgust for a situation.

The correct use of facial expressions is a crucial aspect of social interactions. Studies have shown that smiles can have an impact on how an apology is perceived in court [Pease and Pease, 2008]. Generally, smiling during social interaction allows one to be perceived as more appealing and less threatening.

2.1.2 Oculesics

Oculesics is a form of nonverbal communication and comprises all behaviours related to the human eyes. There are four types of eye behaviours: eye contact, gaze direction, eye movements and pupil dilation. Eye contact plays an important role during social interactions. On the part of the speakers, eye contact is used to regulate turn taking [Kendon, 1990]. Kendon found that when starting a turn, the speaker would look away and make eye contact again when ending the turn, signalling that the listener can now become the speaker. On the part of the listener, eye contact can signal engagement or disengagement. This is why, a standard advice persons receive when preparing for critical interactions such as job interviews is to maintain eye contact. However, Argyle and Cook [1976] found that maintaining the correct amount of eye contact during social interaction is a complex endeavour. Making too much or too less eye contact can make us feel ill at ease [Pease and Pease, 2008].

Related to eye contact is gaze direction. Gaze direction is often used during social interactions to send and perceive feedback [Yngve, 1970], signal points of interest in the physical world [Argyle and Cook, 1976] and to repair the common ground by solving ambiguous verbal or nonverbal instruction [Oviatt, 2003] (e.g. saying "give me that toy" and

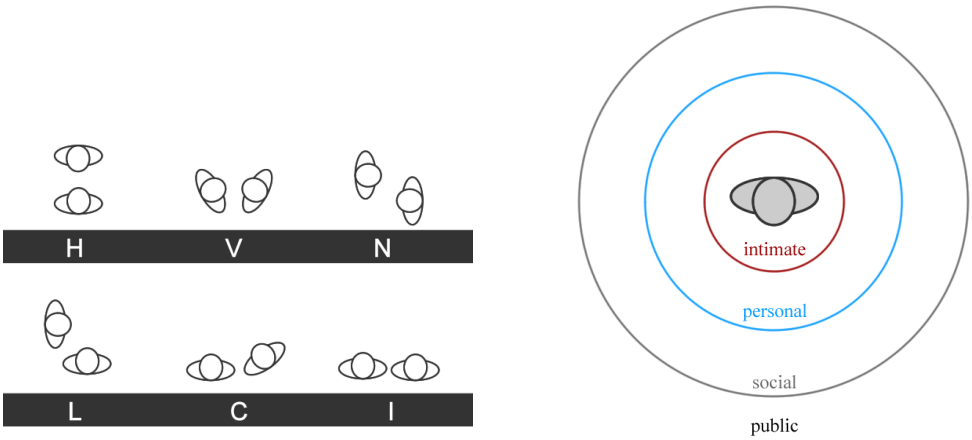

Figure 2.1: Left: Formation variants for a dyadic interaction as described by Ciolek and Kendon [1980]. Right: The four distance zones as defined by Hall [1966].

then gazing towards a specific toy).

The eye can also be used to display emblems. Such eye movements hold specific meaning similarly to arm gestures. For example, rolling one's eye can signal disbelief, boredom or frustration [Armstrong and Wagner, 2003].

Finally, whereas the previous types of ocular behaviour were to some degree voluntary, or at least could be controlled consciously, pupillary dilation is a purely involuntary physiological reaction. Studies have found the size of the pupil to correlate with cognitive load [Kahneman et al., 1969], and thus holds information regarding one's mental state.

2.1.3 Proxemics

A group of individuals that have engaged in an interaction and have agreed to share a common space, enter what sociologists call a formation. According to Kendon [1990], a formation "arises when two or more people cooperate together to maintain a space between them to which they all have direct and exclusive access" and inside which they will conduct the activity that is the scope of the interaction.

Members of a formation may form *arrangements* in space. Kendon names several types of arrangements that can occur in formations: vis-a-vis (H-shaped), L-shaped, side-by-side (I-shaped) and V-shaped. In larger formations we may see linear, circular, rectangular or semicircular arrangements. Ciolek and Kendon [1980] studied the effect specific orientations have on the formation and how others perceive it. They characterize a formation as being either open or closed based on the orientation of the members. The N, H and V-shaped formations are closed while L, C and I are open (Figure 2.1 left). They postulate that humans are more inclined to join open formations as opposed to closed formations.

Another important characteristic of an interaction is the interpersonal distance. Hall [1966] notices that the spacing between individuals in a formation is strongly related to the type of interaction that is happening. He also suggests that the behaviour as well as the perception of the behaviour of others is relative to the distance between the interactants. This distance is defined as a series of circular zones that surround each individual. A visualization of this concept can be seen in Figure 2.1 (right). Hall distinguishes four different distance zones: intimate distance, personal distance, social distance and public distance. The size of each

zone is culture specific, however every culture acknowledges these four zones. For instance, the values for Northern Americans are: intimate distance up to 0.45 m, personal distance from 0.45 m to 1.2 m, social distance from 1.2 m to 3.6 m and the public distance starts at 3.6 m. These values change from culture to culture and what is normal for one cultural group can be considered crowded for another [Ferraro, 1990; Barnlund, 1975]. For example, the average conversational distance for European Americans is approximately 0.5 m. For Latin Americans this distance can drop down to 0.35 m and members of Arabian cultures perceive the conversational distance to be as low as 0.22 m.

2.1.4 Haptics

Nonverbal communication using the sense of touch is referred to as haptic communication. Touch plays an important role in social interactions and is particularly relevant in interactions between persons who have a close relationship. Jones and Yarbrough [1985] identified 7 types of touch: positive affect — used to communicate positive emotions such as support, appreciation, inclusion, sexual intention or general affection; playful — affectionate or aggressive touches intended to diminish the impact of an accompanying verbal or nonverbal message; control — directs behaviour of interlocutor; ritualistic — greeting or departure touches; hybrid touches — a mix of two or more touch types; task-related touches — directly related to a specific task; accidental touches — unintentional touching.

2.1.5 Paralanguage

Paralinguistics, an often overlooked element of nonverbal communication, refers to the characteristics of a vocal message[1]. Thus, there is no utterance which does not have paralinguistic elements. In simpler terms, paralinguistics refers to how we say something, rather than what we say. Typical paralinguistic properties are pitch, intensity, speech rate, utterance duration and others.

Paralinguistic properties can be controlled either voluntarily or involuntarily. For example, we use paralinguistic properties consciously to differentiate between a question and a statement, to signal the use of irony or sarcasm, or to communicate emphasis. However, our physiological or psychological state can also alter the paralinguistic properties of a message. Scherer [2003] postulated that the affective state of the speaker causes changes in "respiration, phonation, and articulation." Other factors which affect paralanguage include social constrains, cultural habits and transmission channel qualities (e.g. echo).

Paralinguistic behaviour plays a critical role during social interactions. For example, in a job interview context, fluency of voice and loudness have been found to correlate with employment decisions [Hollandsworth et al., 1979]. Misuse of paralinguistic behaviour (e.g. due to cultural differences) can also cause misinterpretations of intent or judgement [Gumperz, 1982].

2.2 Human Attention

The social augmentation concept introduced in this dissertation is a typical multi-tasking scenario: the user is asked to perceive, interpret and react to the feedback delivered by the

[1] There are also paralinguistic properties for written and sign language, yet for the purpose of this thesis, the term will be used to describe the properties of the vocal channel

social augmentation system while being engaged in a social interaction. To get a better grasp on how humans manage attention in such situations, we look at psychology to understand the various processes which govern the distribution of attention among competing tasks.

One of the very first definition of attention was postulated by William James in 1980:

> "Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others." (William James [James, 1890])

What William James describes is now referred to as selective attention and is akin to a flash light which can shine on only one mental process at a time. This one "selected" process becomes active whereas the ones left in the dark remain inactive, dormant. To this end, tasks can only be executed in sequence, and any form of "parallelism" is explained as quickly switching attention between processes. A competing view on attention has been proposed in the second part of the 20th century and states that humans have the ability to distribute attention among competing tasks. According to distributive attention models, tasks can be carried out in parallel as long as a "processing capacity" is not exceeded.

In this context, a task is defined as a general cognitive process which needs to be executed to reach a desired outcome. For example, a stimulus (i.e. an environmental event which is within the area of effect of a person's sensory system) is processed by a perception task, upon completion of which the stimulus is perceived (i.e. organized, identified, and interpreted [Schacter et al., 2011]).

The laws of attention are crucial for the work presented in this dissertation as they allow one to understand what effect social augmentation feedback (a secondary task) can have on the social interaction (the primary task). To this end, this section will provide a quick introduction into the human attention mechanism. It will present the most popular attention models and discuss how they compare to each other. Furthermore, the interaction between multiple tasks and how a task's attention requirements can change over time will also be discussed.

2.2.1 Cognitive Models of Attention

One of the first cognitive models of attention introduced was Broadbent's bottleneck model [Broadbent, 1957] which states that at every point in time only one task can be processed or attended to (see Figure 2.2 left). Thus, the model is in line with William James' view on attention. Imagine yourself at a party talking to a person while other persons also have conversations in the same room. Do you know what the other persons are talking about? Most of time the answer is no, because according to the bottleneck model, we can only process a single stimulus (i.e. the audio signal of one conversation) at the same time.

A different type of model, which complements the bottleneck model rather than competing with it, is the capacitive model [Kahneman, 1973; Moray, 1967]. It states that at any given point in time we only have a limited capacity to dedicate to processing tasks. This capacity can be split up between activities as long as there is free capacity left. Figure 2.2 (right) provides an illustration of how a capacitive model extends the bottleneck concept by more precisely defining the bottleneck.

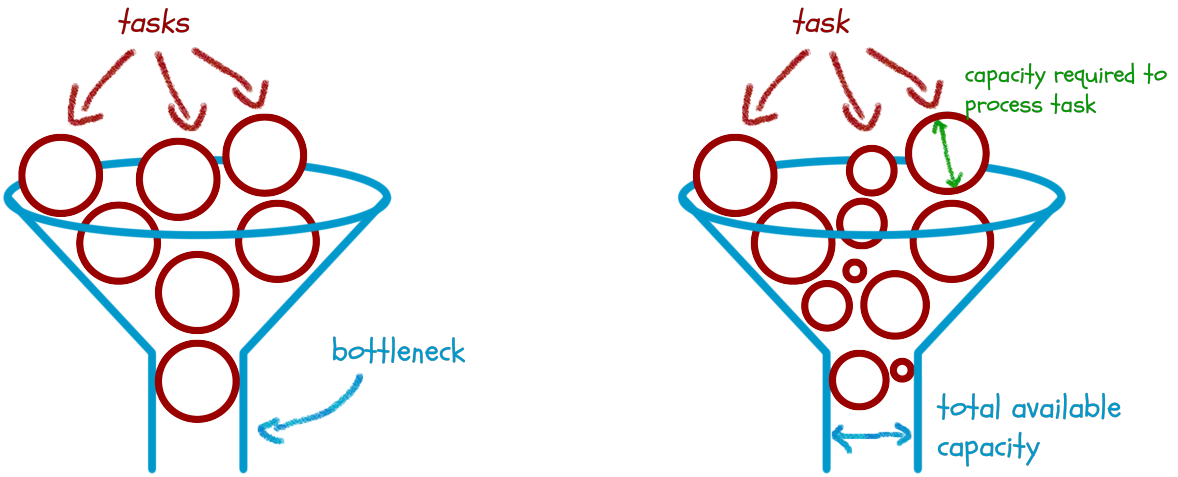

Figure 2.2: Illustrations of bottleneck (left) and capacitive models (right) of attention.

According to the capacitive model, each task requires a certain amount of resources for processing. Kahneman et al. [1968] claim that the capacity demanded by an activity is proportional to its difficulty. In the case that there is not enough capacity available for a given activity, either because the total amount of capacity is insufficient or because not enough is allocated to this activity, task execution is either suspended or task performance decreases. For example, when driving on a straight road we can usually have a conversation with our passenger without much trouble. However, as soon as we arrive at a crossroads we engage in a more difficult activity during which we pause the conversation.

Since its proposal, Kahneman's capacitive model has been the subject of much discussion in the field. One common criticism regarded interferences between modalities. It is obvious that tasks which belong to different modalities are easier to execute simultaneously than tasks belonging to the same modality. For example, I have no problem listening to music while cooking but I usually turn the music off when I talk to someone else. Kahneman agreed that such effects cannot be explained by his capacitive model and introduced the concept of structural interferences [Kahneman et al., 1968]. He theorized that it is easier for man to divide attention between tasks which do not compete for the same cognitive mechanisms. In the example above, it is easier to cook (visual) and listen to music (auditory) than to converse (auditory) and listen to music (also auditory). Brooks [1968] tested this effect in various experiments. In one, users were asked to recall a sentence (e.g. "A bird in the hand is not in the bush") and to specify for each word whether it is as a noun or not by either tapping with the foot, the hand or by saying "yes" or "no." The experiment found the verbal task to be far more difficult for the users to perform than the tapping tasks.

Whereas for Kahneman, structural interferences where just the exception to the rule, Navon and Gopher [1979] more explicitly addressed this issue by accounting for multiple resource pools. According to them, tasks demand multiple types of resources which can be drained from different resource pools. Each task demands a specific combination of resources. Depending on the availability of resources, multiple tasks can be performed in parallel without performance degradation. In the example from Figure 2.3, only task 1 and 3 can be executed simultaneously without performance degradation due to a shortage of circle and triangle resources. Forcing task 1 and 2, 2 and 3 or all three to run in parallel will most likely severely impact the performance of one or more tasks. This model explains some empirical

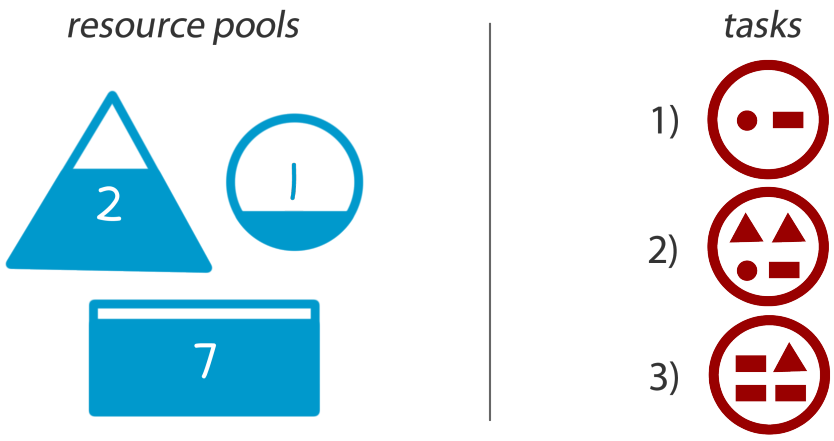

Figure 2.3: A multiple-resource model of attention as proposed by Navon and Gopher.

findings which Kahneman's global resource pool cannot. For example, Wickens et al. [1977] (referenced in [Navon and Gopher, 1979]) observed that the performance of a primary task was impacted similarly regardless of whether the user had to perform a simple or a more difficult secondary task. This "difficulty insensitivity" [Wickens, 2008] suggests that the more difficult secondary task required additional resources which were irrelevant to the primary task. While this effect can be also be labelled as a structural interference [Kahneman et al., 1968], it "seems inadequate once we realize that processes which use the same mechanisms sometimes interfere with each other but they seldom block each other completely" [Navon and Gopher, 1979].

A refinement (and specialization) of Navon and Gopher's mathematical model has been proposed by Wickens [Wickens, 2002]. The model can be illustrated as a cube (Figure 2.4) with four dimensions: stages of processing (perception, cognition and responding), codes of processing (spatial and verbal), modalities (visual and auditory) and visual channels (focal versus ambient vision). The general idea of the model is simple: As long as "two tasks use different levels along each of the [four] dimensions, timesharing will be better." [Wickens, 2008]. The model has found great popularity in the automotive field for predicting task performance in driving scenarios.

Initial theories of selective attention (e.g. the bottleneck model [Broadbent, 1957]) proposed that only attended stimuli are analysed and that the others are discarded without processing. Other researchers however, most notably Deutsch and Deutsch [1963] and Trumbo and Noble [1970], contradicted this view, suggesting the bottleneck comes only later. According to them, all stimuli are perceived and processed, but we only respond to some of them. In the party example from the beginning of this section, this would mean that we do hear the other conversation in the room but we only respond to (i.e. rationalize, memorize) one conversation. This view explains why, in such a situation, we can still react if someone calls our name. This theory is compatible with the concept of subliminal perception, i.e. man's ability to perceive and extract information from stimuli without the use of conscious processes (see Section 2.2.3). Modern models are more flexible and thus moved past this debate. Wicken's four dimensional attention cube [Wickens, 2002] and Navon and Gopher's resource model [Navon and Gopher, 1979] more generally explain the information processing mechanism and postulate that resources are consumed at every processing stage. This means

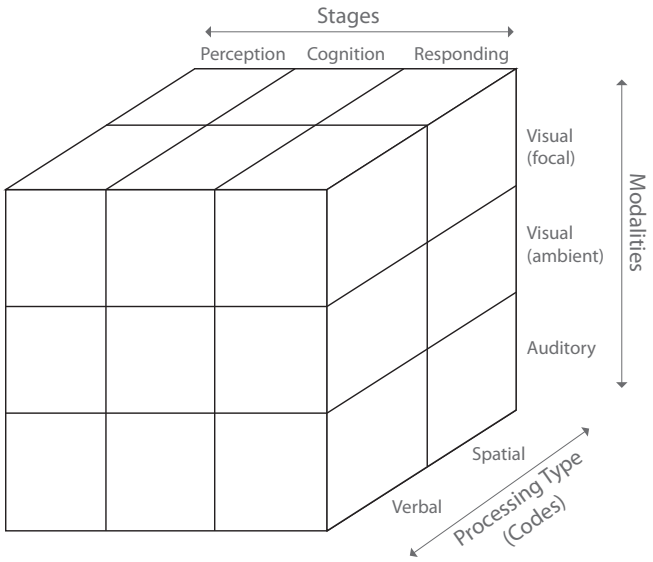

Figure 2.4: An illustration of the multiple resource model as proposed by Wickens [Wickens, 2002]. For easier understanding, Wickens' fourth dimension ("visual processing") has been represented on the "modalities" axis, transforming the 4D cube into a 3D one.

that, the bottleneck can occur everywhere there is an insufficiency of resources and thus explains both variants.

To sum up, for two tasks to be executed in parallel, the existence of enough resources for satisfying the demand of both tasks is required. Thus, simple (i.e. which demand few resources) and dissimilar tasks (i.e. which demand different tasks) are more likely to be executable in parallel. For our social augmentation scenario we can conclude that, simple feedback mechanisms which use modalities different from those of the primary task, are more suitable for parallel execution, and thus are less likely to disrupt with the primary task.

2.2.2 Task Interruption

In the previous subsection we looked at how humans can process multiple tasks at the same time and what conditions need to be met to enable this. However, one question remains open: What happens if two tasks cannot be processed in parallel, and one task interrupts the other?

Researchers distinguish two types of interruptions: internal or external [Miyata and Norman, 1986]. Internal interruptions are caused by voluntary (also called top-down) switches in attention. These are generally goal-oriented, i.e. the user consciously diverts attention away from the primary task to satisfy a particular goal (e.g. extract information). On the other hand, external interruptions are caused by stimulus-driven, involuntary attentional capture, i.e. the attentional shift happens due to a characteristic of the stimulus rather than the intention of the person. For example, sudden movement in our field of view will involuntarily capture our attention. The most common property of a stimulus for causing involuntary attentional capture is novelty [Johnston et al., 1990; Remington et al., 1992]. More precisely, new objects in our field of view tend to "pop out", automatically capturing our attention. Johnston et al. [1990] describes this effect as the "automatic orientation of attention away from more fluently unfolding regions of the perceptual field (familiar objects) and toward less fluently unfolding regions (novel objects)." Besides novelty, researchers also found that involuntary attentional

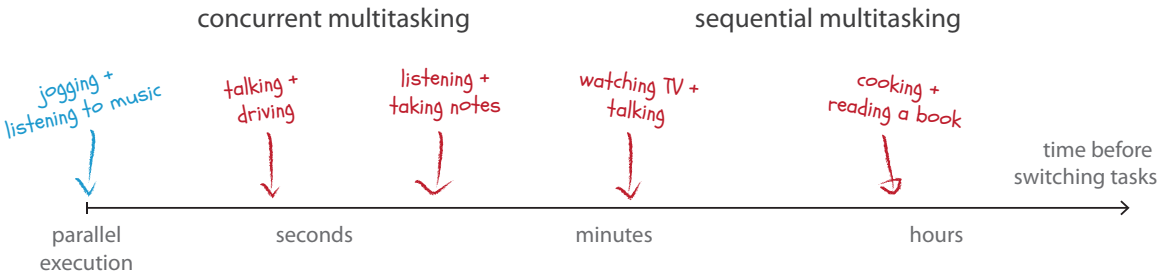

Figure 2.5: The multitasking continuum, adapted from Salvucci et al. [2009].

capture prioritizes value-infused [Anderson et al., 2011] or emotional stimuli [Notebaert et al., 2011], as well as regularities in stimulus presentation [Zhao et al., 2013]. Moreover, not all novel stimuli may capture attention. Perceptual studies suggest that the saliency of a stimulus correlates with the probability of triggering an attentional shift [Berlyne, 1960; Theeuwes, 2010; Zehetleitner et al., 2013].

For social augmentation, both types of attentional capture are of interest. If the user has the necessary cognitive resources to actively seek improvement of their behaviour, voluntary attentional capture will be the predominating method for extracting information from feedback. However, if the primary task is too demanding, it is likely that the user will not actively pursue feedback perception. In this case, only involuntary attention capture may cause feedback perception.

A situation in which a user processes multiple tasks over a given time window is commonly referred to as multitasking. Salvucci et al. [2009] introduced the *Multitasking Continuum* to classify such situations according to the amount of time elapsed between two consecutive switches. Their continuum ranges from concurrent multitasking situations, rapidly switching between tasks with only few seconds elapsing between switches (e.g. driving and talking), to sequential multitasking where minutes or even hours can pass before a switch happens (e.g. cooking and reading). Interestingly, in their original paper, Salvucci and colleagues do not explicitly include fully parallel task processing (as discussed in the previous section) in the *Multitasking Continuum*, leaving the impression that in all multitasking situations, tasks are processed sequentially. Nevertheless, it is plausible to think about divided attention-based parallel processing as being at the very beginning of the continuum (see Figure 2.5).

One measure researchers often use to assess the impact of a secondary task on the primary task is task switching-induced lag [Altmann and Trafton, 2004]. There are two types of lag: interruption and resumption lag. The interruption lag is defined as the time which passes between the moment of interruption and when the actual processing of the secondary (interruptive) task starts. Similarly, the resumption lag is defined as the time it takes between the end of the secondary task's processing and the time it takes to restart the (previously suspended) primary task [Altmann and Trafton, 2004]. In both cases, the delay is thought to be caused by "mental housekeeping", such as freeing up and allocating cognitive resources. However, literature suggests that this delay can be reduced with the acquisition of time-sharing skills [Wickens and McCarley, 2007], which make the user more apt at allocating resources and interrupting and resuming processes.

When designing for multitasking situations such as our social augmentation concept, paying close attention to the quality and quantity of interruptions is crucial. Empirical

findings show that interruptions (internal or external) can have a large impact on overall performance, with even short interruptions doubling the number of errors a person makes during an abstract character analysis task [Altmann et al., 2014]. Studies have also shown that the moment when the interruption happens relative to the state of the primary task impacts the size of the interruption and resumption lag. More specifically, if the interruption happens at moments of low cognitive demand, such as subtask boundaries, the resulting interruption and resumption lag is shorter [Bailey and Iqbal, 2008]. Furthermore, there is evidence that the nature of the interruption is also important. For example, content-poor interruptive tasks (e.g. blank screens [Monk et al., 2004]) and tasks with related content [Cutrell et al., 2000] are less disruptive than content-rich and unrelated tasks. Starting from these findings, it is clear that in our scenario, it is critical to control the amount, complexity, duration and timing of live feedback to lessen the impact on the primary task.

2.2.3 Unconscious Perception of Stimuli

Unconscious cognition refers to any process which is being executed by our brain without us being aware of it. For example, we are able to carry out certain tasks (e.g. walking, driving) or respond to certain stimuli (e.g. turning the head when someone calls our name, scratching our skin when it itches) without consciously deciding to do so. While there are multiple (sometimes conflicting) definitions for conscious cognition [Dehaene et al., 2006; Greenwald, 1992], for the scope of this manuscript the simple definition given above suffices.

One subtype of unconscious cognition is unconscious perception of stimuli. Stimuli which are perceived unconsciously are called subliminal, a term formed from the Latin word for *below* ("sub") and *threshold* ("limen"). Thus, a subliminal stimulus is an external stimulus which, upon sensory reception, is processed unconsciously, i.e. without the receptor being aware of it.

Starting from the models of attention described in the previous section, it is plausible to state that subliminal stimuli are simply stimuli which, for some reason, do not reach the final stage of processing (i.e. response) and thus consciousness. However, this view suggests that subliminal stimuli are simply discarded somewhere on the way and thus are of no particular significance. Empirical findings contradict this view. One of the most successfully replicated studies, and also the most controversial, involve the *Subliminal Psychodynamic Activation* (SPA) effect. In one such study type, participants are shown a visual stimulus containing the text "Mommy and I are One," typically for 4 ms. This subliminal presentation has been shown to have significant effect on the participants' behaviour, helping individuals quit smoking and even yielding therapeutic gains [Bornstein, 1990; Silverman and Weinberger, 1985; Weinberger, 1992]. This effect can be explained by a strong unconscious bond with the "mother of early childhood" resulting in a drive for improvement upon stimulation [Silverman and Weinberger, 1985]. In a different line of experiments it has been shown that while actively focusing on one input channel, we can unconsciously perceive characteristics of stimuli from an unattended channel. For example, in a dichotic listening study, Cherry [1953] has demonstrated that while attending to audio messages on one ear, participants were still able to recall low-level features such as pitch, loudness and spatial locations of sounds played on the other ear.

These findings suggest that subliminal stimuli do in fact reach the end of the information processing pipeline and trigger a response, but do so in a way which does not involve the human consciousness. One of the first researchers to study this effect was Greenwald [1992].

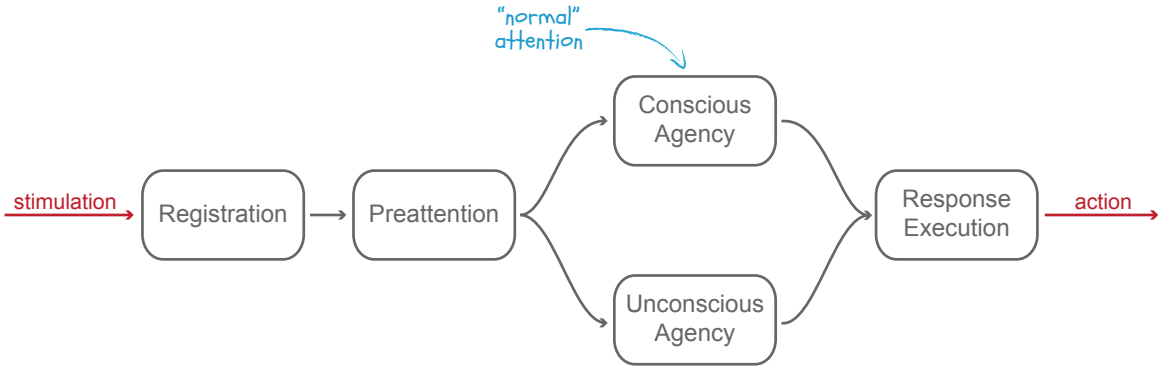

Figure 2.6: Information processing pipeline explicitly modelling conscious and unconscious processing paths, adapted from [Greenwald, 1992].

He addresses it by proposing a model which explicitly includes unconscious perception as part of the information processing pipeline, which constitutes an alternative processing path to conscious perception (see Figure 2.6).

As discussed in Section 2.2.1, concurrent tasks compete for cognitive resources and, if not enough resources are available, task performance decreases or even breaks down. Considering this, the prospect of completely delegating some tasks to the unconscious and freeing up valuable resources is particularly enticing. However, here it is important to point out that unconscious is not the same as effortless. According to resource-based attention theories [Wickens, 2002], every stage of the processing pipeline consumes resources and it is plausible to assume that this rule extends to unconscious processing phases as well. However, since unconscious processing does not use the same processing stages as conscious processing [Greenwald, 1992], "conscious resources" are free to be used elsewhere.

Providing the user with feedback which can be processed and reacted to unconsciously promises a seamless augmentation of the primary task. Such subliminal feedback is composed of stimuli which are designed to only be processed unconsciously, and thus fall under the threshold of conscious perception. This could be achieved by modulating the characteristics of these stimuli in a way that they will always go on the unconscious processing path. Typical examples of such stimuli are: visual information which is presented for a very short time (literature suggests a limen of 4 ms [Hardaway, 1990]) or with reduced brightness [Baker, 1937]; auditory stimuli presented at less-than-perceptible intensities [Baker, 1938]; or subtle changes in physical properties of objects, such as weight [Pierce and Jastrow, 1984].

2.2.4 Task Automation

From our day to day lives we know that tasks become easier with time. For example, I recently started driving again after an eight year break and I found the complexity of the task overwhelming at first. Driving was difficult, highly attention demanding and stressful. In this initial period, I was utterly unable to do anything else except driving: The radio was always turned off and conversations were strictly prohibited for all passengers. However, the difficulty of the task and its thirst for attention seemingly diminished over time. Within weeks, I started using the radio, after a few months I was able to have simple conversations with other passengers and now, over two years later, I am able to hold political debates in the car without feeling overloaded. Navon and Gopher [1979] addressed this dynamic nature of

cognitive resource demand in their economy-inspired model of attention. According to them, task practice can have three possible effects: (1) the cognitive system learns how to utilize the resources demanded by the task more efficiently; (2) practice reduces the resource demand of the task; and (3) practice reduces the organisational overhead associated with resource allocation.

After lengthy practice, some tasks may become automated to a degree that they no longer require "active control or attention by the subject" [Schneider and Shiffrin, 1977]. Schneider and Chein [2003] characterize automated tasks as fast and parallel, requiring little effort and being "robust to stressors" such as alcohol, fatigue, stress or vigilance. Studies have also shown that automated tasks are executed with high accuracy and low rate of error [Altmann et al., 2014]. However, to achieve automation, "extended consistent training is required," and once a task becomes automated, it can be difficult to control. Thus, there are both advantages and disadvantages to automation.

In a social augmentation scenario, automating the reaction of the user to feedback can benefit the effectiveness of the augmentation. More specifically, the reduced strain on the cognitive mechanisms means that feedback perception and interpretation can easily be carried out in parallel to the social interaction. Furthermore, automated tasks are also activated automatically [Schneider and Shiffrin, 1977] in response to specific stimuli, making not only the processing of the task but also the decision making prior to the execution unconscious, automated and less cognitively demanding.

2.3 Summary

This chapter provided the reader with an overview of the theoretical background for the concept of social augmentation. First, Section 2.1 looked at social interactions in general and discussed the role of the different social behaviours in an interaction. Moreover, it illustrated what happens when and if social behaviours are used inappropriately. In such situations, social augmentation systems could be used to inform the users of how others perceive their behaviour, and it can be adjusted to minimize the likelihood of misunderstandings and generally achieve a better interaction outcome.

The second section looked at the process behind feedback perception. For this, an overview of literature from cognitive psychology has been provided to create a basic understanding of how humans distribute attention across multiple tasks, and motivate the need for careful deliberations when it comes to delivering live feedback to users engaged in social interactions.

The ideas and theories introduced in this chapter will be used in the remaining of the dissertation as a foundation for the design of the social augmentation concept.

3. Social Skills Training

Social skills training represents the systematic teaching of specific behaviours to help a person develop or improve the effectiveness of their interpersonal interactions [Goldstein et al., 2013]. Over the past century, it has been used both in an educational context to help individuals become better leaders, manage stress, make a better impression during job interviews or deliver more compelling public speeches, but also in a medical context to aid persons suffering from mental disabilities better cope with day-to-day life.

A variety of methodologies have been developed for training social behaviour. Classical approaches involve the learner memorizing certain behaviour patterns from different media sources. The learned behaviour patterns are then usually evaluated by a coach who provides feedback on how well the knowledge was assimilated. Following this, a new iteration of the learning process starts in which the learner applies the received feedback to further improve their skills. More recently, technological progress has enabled the development of automatic and semi-automatic instruments which offer easier, more effective and more accessible forms of training. Such technology-enhanced training tools can record the social behaviour of a learner, which can then be subsequently inspected by the learner or a coach for problematic behavioural aspects. Through repetition, the learners can then work on the identified problems and ultimately improve their general skill. Other, more complex systems make use of simulation techniques to transport the user in virtual environments where they can interact with virtual agents as if they would with real persons. Popular scenarios for such training systems include job interviews [Anderson et al., 2013; Baur et al., 2013a; Hoque et al., 2013], public speaking [Batrinca et al., 2013; Chollet et al., 2015] or social anxiety training [Pan et al., 2012].

To illustrate the current state-of-the-art in terms of technology-enhanced social skill training, this chapter will first introduce one example of such a system, aimed at teaching pupils job interview-pertinent social skills. Following this, the chapter will reflect upon the limitations of such systems, and make a case for the need to switch to a more flexible and direct form of improving one's social behaviour.

Figure 3.1: The interface of TARDIS showing the virtual character playing the role of an interviewer.

3.1 Job Interview Training

Compared to classical learning approaches (e.g. coaching), technology-enhanced solutions such as serious games present themselves as viable and advantageous alternatives [Stapleton and Taylor, 2003]. Their automated nature gives users access to personalized feedback without the need for human coaches, improving scalability and repeatability. This reduces the running costs of such systems and makes them a viable solution for mass deployment.

From a recruiter's point of view, the goal of a job interview is to determine the fit of the candidate to a particular position in the company by evaluating the candidate's verbal (i.e. content of utterance) and nonverbal behaviour (e.g. use of voice, gestures, postures, facial expressions). Nonverbal behaviour is particularly critical as research shows it takes a significant role during interpersonal interaction [Birdwhistell, 2011; Mehrabian, 1981]. In particular, studies [Carl, 1980; Hollandsworth et al., 1979] show that nonverbal behaviour has a large impact on the outcome of a job interview. Training such behaviour can therefore be very beneficial to improving one's chances for employment.

This section[1] introduces the TARDIS virtual job interview game for training young adults. The system allows users to take part in gamified job interviews led by a virtual character. Using social signal processing techniques, the system records and analyses the user's nonverbal behaviours, which are then used to trigger actions for the virtual characters in realtime, but also as material for a semi-automatic debriefing phase.

The section will first describe the job interview training system. Following this, a three-day user study with 20 pupils is presented. The aim of the study was to measure the impact of the system on the pupils' job interview performance and compare it to a conventional learning method commonly used by the school (learning from a written job interview guide).

3.1.1 The System

TARDIS is a serious game designed to help young adults at risk of exclusion explore, practise and improve social skills pertinent to job interviews. The system is composed out of a real time

[1] This Section is an adaptation of Damian et al. [2015a]

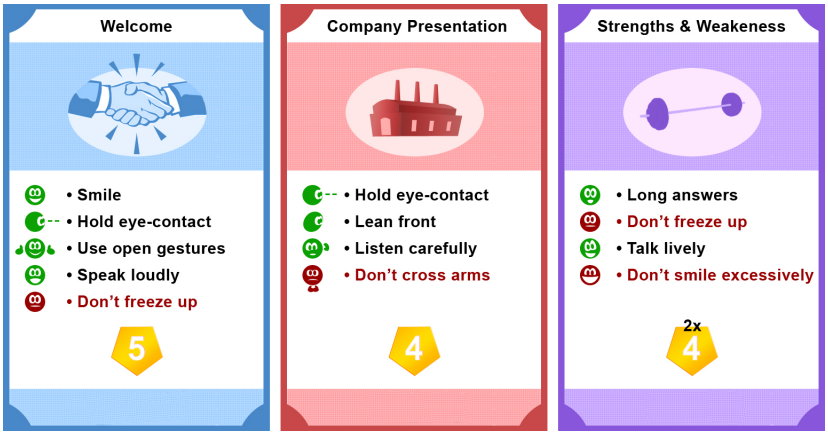

Figure 3.2: The game cards give hints to the pupils regarding appropriate behaviour for upcoming interview phases (right).

behaviour recognition system, a scenario manager, and the Charamel 3D character rendering environment[2]. The embodiment of the system is a virtual character able to autonomously conduct a job interview with the user by displaying both verbal and nonverbal behaviour (Figure 3.1 left).

TARDIS makes use of the Social Signal Interpretation (SSI) framework [Wagner et al., 2013] for performing realtime behaviour analysis. More precisely, the behaviour of the user is first captured using a Microsoft Kinect[3] sensor and then analysed in realtime with the help of a processing pipeline. During this process, various behavioural cues, such as gestures, postures or speech characteristics, are extracted and forwarded to the scenario manager. Based on the user's behaviour, the scenario manager [Gebhard et al., 2012, 2014] chooses an appropriate reaction for the virtual character and forwards it to the rendering engine. Besides impacting the behaviour of the virtual character, the results of the behaviour analysis are also stored for later use during a post-hoc debriefing phase.

Two job interview scenarios have been implemented using the VisualSceneMaker authoring tool [Gebhard et al., 2012]. Both scenarios have been developed in cooperation with practitioners and chosen for their relevance to our target user group. The first scenario simulates a job interview for a position as an electro-mechanical engineer and is aimed at male pupils. The second scenario has been designed for female pupils and represents a job interview for a trained retail salesman position. Both scenarios are structured in three phases (*Welcome*, *Company Presentation*, and *Strengths and Weaknesses*). Each phase contains several turns during which the virtual character will ask the user a question and then wait until an answer is given.

In an effort to keep the pupils engaged and motivated, various game-like elements have been added to the system. More precisely, we introduced physical game cards which are similar in appearance to those of classic board games and give hints on how to behave during the interview (Figure 3.2). A scoring system keeps track of how well a pupil follows these hints. For example, if the pupil smiles at an appropriate moment, the score will get incremented by one (Figure 3.1).

After the simulated interview, the users take part in a semi-automatic debriefing phase.

[2] `http://www.charamel.com`
[3] `https://developer.microsoft.com/en-us/windows/kinect`

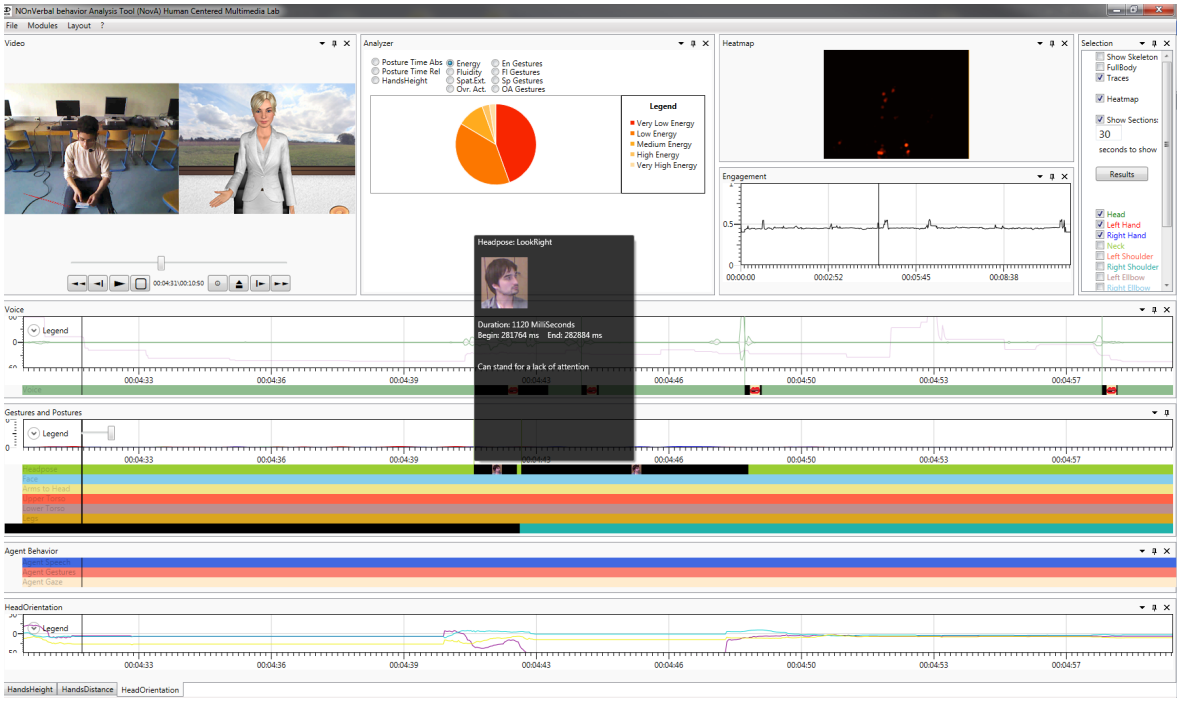

Figure 3.3: The interface of the NovA debriefing tool.

For this, the system makes use of the NovA analysis tool [Baur et al., 2013b]. NovA places the recorded behaviour of the user on the interview time line (see Figure 3.3) and allows users to inspect and reflect upon their performance during the interview. Important events are highlighted, and an interpretation of their significance for the job interview is provided in text form.

3.1.2 Evaluation

To evaluate the effectiveness of the TARDIS system at teaching job interview-pertinent social skills, an evaluation was conducted in cooperation with a school in Stadtbergen/Germany. The objective was twofold: On the one hand we wanted to investigate whether pupils' job interview skills are rated better by practitioners after using the TARDIS interactive job training game compared to before. On the other hand, we wanted to evaluate whether their skills are rated better, or at least equally well, in comparison with pupils who trained using conventional teaching methods.

Participants

In total, 20 pupils (10 male and 10 female) from the eighth and ninth grade (final and second to last school years) have been recruited to take part in the study. Participants were aged between 13 and 16 (mean = 14.37; SD = 0.94). The data of one participant had to be removed due to extraordinary circumstances resulting in nervous and unfocused behaviour (she accompanied her friend to the hospital after a minor accident).

Additionally, two career counsellors participated in the study as professional practitioners. The career counsellors are employed full time at Career Service - Augsburg University, where they advise students on choosing suitable jobs, preparing their application documents, and training for job interviews.

	experimental group	control group
day 1	mock job interview	mock job interview
day 2	interaction with training system	training with book
day 3	mock job interview	mock job interview

Table 3.1: Procedure of user study over three days.

Figure 3.4: Mock job interview with professional career trainers (left) and interaction with job interview training system (right).

Procedure and Apparatus

The user study was conducted over the course of three days. An overview of the procedure can be seen in Table 3.1. On the first day, all pupils participated in mock job interviews led by a practitioner (see Figure 3.4 (left)). The purpose of these mock interviews on the first day was to establish a baseline regarding the job interview performance of the pupils prior to training. Furthermore, as the system's goal is to help the users improve their nonverbal behaviour, the practitioners were also asked to focus on the nonverbal behaviour, i.e. how the participants answer rather than what they say. Two interviews were carried out in parallel in separate rooms whilst each lasted for approximately 7 minutes. This duration was deemed sufficient by the practitioners to get an objective measurement of the pupils' job interview performance. After each mock interview, both pupils and practitioners filled out questionnaires *A* and *B* respectively.

In *Questionnaire A*, practitioners rated 1) the pupil's overall performance, 2) whether they would recommend the pupil for employment, 3) appropriate usage of smiles, 4) appropriate usage of eye contact, 5) appropriate usage of gestures, as well as whether the pupil seemed 6) nervous 7) interested and 8) focused. In *Questionnaire B*, pupils self-reported on whether they thought they 1) performed well in the interview, 2) were nervous, 3) used a lot of filler words such as "er" or "uhm", 3) were focused, 4) were aware of their non-verbal behaviour and 5) performed appropriate non-verbal behaviour. Both questionnaires used Likert scales ranging from 1 to 7, with a higher value indicating a better performance. The only exception is the nervousness dimension, where a lower score is considered being better.

On the second day, pupils were randomly divided into experimental group (EG) and control group (CG), resulting in four females and six males for the EG and five females and four males for the CG.

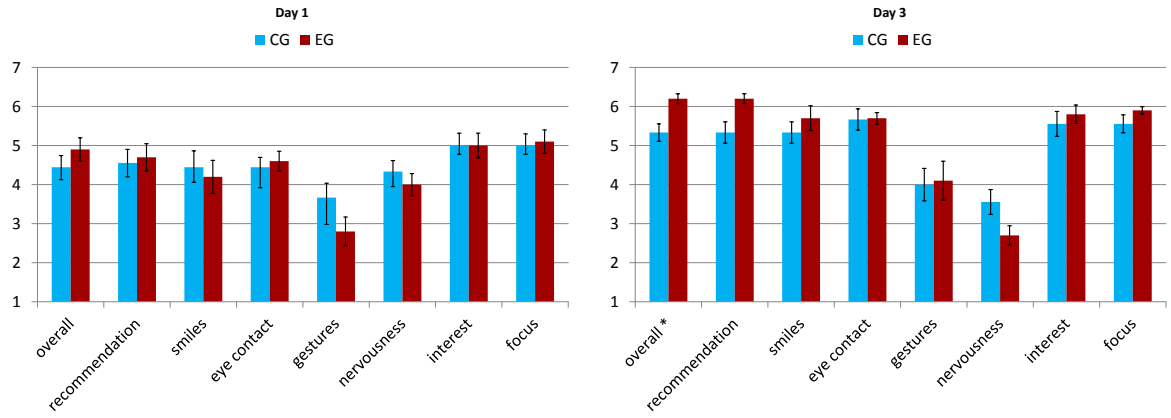

Figure 3.5: Practitioners' ratings of day one (left) and day three (right) comparing CG and EG. Dimensions marked with * present significant differences between the two groups.

The EG interacted with the TARDIS system. Figure 3.4 (right) shows a sample interaction with the system in which a pupil analyses one of the game cards. The participant was seated at a school desk with a Microsoft Kinect and a webcam positioned to face them. During the interaction, the participant was also wearing a close-talk microphone. Each training lasted for about 15 minutes, split between game interaction and debriefing. During the training session, the pupils' nonverbal behaviour was recorded and analysed by the system. In the debriefing phase, a researcher assisted the pupils in reviewing their performance using the NovA analysis tool [Baur et al., 2013b]. However, the researcher only provided technical support with the system and helped the pupils operate the interface.

Pupils of the CG were reading a job interview guide[4] for the same amount of time. The written guide is published by a local youth advisory institution and regularly used by our cooperating school to prepare their pupils for job interviews.

On the third day, a second round of mock job interviews was conducted with each participant. Pupils of both groups (EG and CG) were brought to the practitioners in random order, who were unaware of which condition the pupils have been assigned to during the second day. After each mock interview, pupils and practitioners filled in the same questionnaire they filled in during day one (questionnaires A and B respectively). This allowed us to compare the pupils' performance between day one and three.

Results

To determine the quality of the results we used independent two-tailed t-tests when comparing between groups, and paired two-tailed t-tests when comparing between days. In both cases we apply the Bonferroni-Holm error correction method to adjust the significance levels. Analysing the first day of our experimental setup, no significant differences were found in questionnaires A and B when comparing pupils that were later assigned to either join EG or CG (see Figure 3.5 (left)).

Comparing the two groups again on the third day (after either having used the system or the written guide on the second day) revealed interesting insights. We found statistically significant differences for the practitioners' ratings on overall performance ($p = 0.005$, $\alpha =$

[4] https://bayern.aok-on.de/berufseinsteiger/beruf-zukunft/koerpersprache-im-vorstellungsgespraech

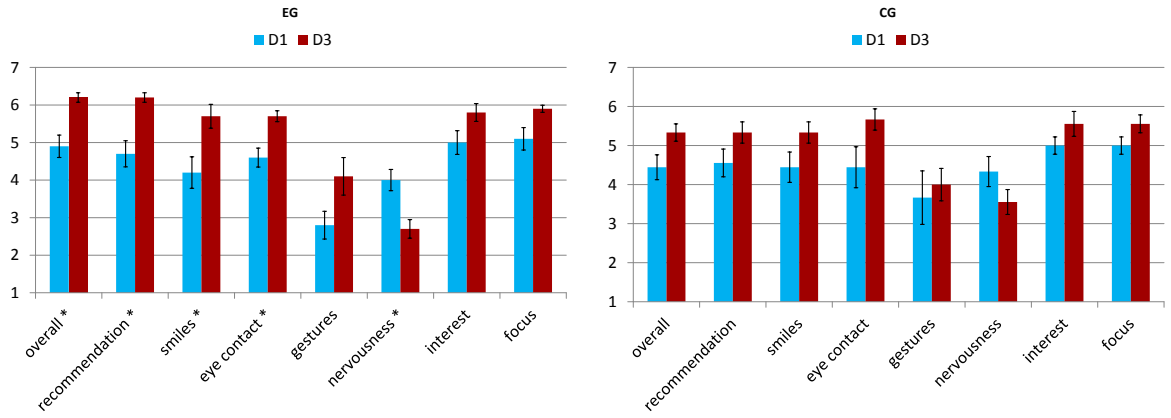

Figure 3.6: Practitioners' ratings of EG (left) and CG (right) across day one and three. Dimensions marked with * present significant differences between the two days.

0.006), with pupils of the EG being rated better compared to pupils of the CG. A strong trend was also found for the recommendation dimension ($p = 0.012$, $\alpha = 0.007$). All other dimensions were also rated better for the EG than for the CG, albeit not significant. Figure 3.5 (right) illustrates these results.

In order to evaluate the improvement of performance for each group individually, we compared the results within groups between day one and three. Our tests revealed significant differences for the EG for the dimensions recommendation ($p = 0.005$, $\alpha = 0.006$), overall performance ($p = 0.006$, $\alpha = 0.007$), nervousness ($p = 0.006$, $\alpha = 0.008$), eye contact ($p = 0.007$, $\alpha = 0.010$) and smiles ($p = 0.012$, $\alpha = 0.013$) (Figure 3.6 (left)). No significant improvements in the practitioners' ratings have been found for the CG when comparing day three to day one (Figure 3.6 (right)). However, trends have been found for smiles ($p = 0.021$, $\alpha = 0.006$), overall performance ($p = 0.035$, $\alpha = 0.007$) and eye contact ($p = 0.047$, $\alpha = 0.007$).

Regarding the pupils' self-assessment, no significant differences were found between the two groups on the first and third day. However on the third day, a strong trend was found for the nervousness dimension ($p = 0.030$, $\alpha = 0.008$), with pupils of the EG rating themselves as less nervous than pupils of the CG. Comparing the two days for each group separately, reveals a significant difference on the nervousness dimension ($p = 0.001$, $\alpha = 0.008$) for the EG only, with participants rating themselves being less nervous on the third day compared to the first day.

Table 3.2 gives an overview of the mean ratings from both questionnaires on the first and on the third day for both conditions.

Discussion

The analysis of the questionnaire data on the first day of our experiment revealed no significant differences, suggesting that the two groups were performing equally well in their job interviews. We can thus consider differences observed on the third day between the groups to be caused by the training completed on the second day. In general, both groups improved from the first day to the third. This is not surprising considering the fact that all participants preoccupied themselves with the topic of job interviews over the course of three days. However, only for the pupils of the EG were we able to observe significant differences

Questionnaire A	day 1		day 3	
	CG	EG	CG	EG
overall performance	4.44	*4.9*‡	**5.33***	***6.2***‡*
recommendation	4.55	*4.7*‡	5.33	*6.2*‡
smiles	4.44	*4.2*‡	5.33	*5.7*‡
eye contact	4.44	*4.6*‡	5.66	*5.7*‡
gestures	3.6	2.8	4.0	4.1
nervousness	4.33	*4.0*‡	3.55	*2.7*‡
interest	5.0	5.0	5.55	5.8
focus	5.0	5.1	5.55	5.9

Questionnaire B	day 1		day 3	
	CG	EG	CG	EG
overall performance	4.66	4.6	5.33	5.2
nervousness	4.77	*4.2*‡	4.33	*2.2*‡
use of filler words	4.88	3.4	4.22	3.0
focus	4.23	3.9	4.56	4.6
aware of n.v. behaviour	5.0	5.0	5.77	5.4
n.v. behaviour	5.22	4.7	5.77	5.1

Table 3.2: Mean values of control group (CG) and experimental group (EG) on first and third day. Significant differences between groups on a particular day are written in bold and marked with *. Significant differences within groups between days are written in italic and marked with ‡.

(for the dimensions overall performance, recommendation for the job, smiles, eye contact and nervousness). Furthermore, when comparing the two groups on the third day, practitioners rated the overall job interview performance of the EG significantly better than that of the CG. This suggests that the technology-enhanced training had a greater effect on the pupils' job interview performance than the traditional method.

The only statistical difference found in the pupils' ratings was the self-reported nervousness of the EG between day one and three. This is also interesting as it indicates that the training system might help users feel more comfortable during job interviews.

The system also left a good impression on the school teachers who stated that "using the system, pupils seem to be highly motivated and able to learn how to improve their behaviour [...] they usually lack such motivation during class." As a possible reason for this, they mentioned the technical nature of the system, which "transports the experience into the youngster's own world" and that the technology-enhanced debriefing phase "makes the feedback be much more believable." Pupils also seemed to enjoy interacting with the system. Most of them asked questions regarding how the score was computed, and which of their behaviours contributed to the final score. This suggests that the scoring functionality had a positive effect on the pupils' engagement in the exercise.

Similar results have been found by other researchers as well. For example, investigations conducted by Pan et al. [2012] suggest that a party simulation involving a virtual female character can help reduce social anxiety in young adult males. Sapouna et al. [2010] studied the effect of a virtual learning system to reduce the bullying victimisation rate of children in schools. The system exposed pupils to simulated bullying situations and allowed them

to experiment with coping strategies. A user study yielded that the system had a positive effect on the children's abilities to cope with bullying. Another relevant work is the user study conducted by Hoque et al. [2013]. They explored the impact of a job interview training environment on MIT students. They conclude that students who used the system to train, experienced a larger performance increase than students who used conventional methods.

3.2 Limitations of Current Approaches

The previous section showed that computer-enhanced social skills systems can help users improve their social skills. However, such training methods still have various limitations which prohibit them from finding large scale adoption by the general public.

One issue is that such simulation environments usually offer a small variety in terms of supported scenarios or possible actions the user can explore. For instance, the TARDIS job interview training system supported only two different interview scenarios: mechanic and saleswoman. While these two scenarios loosely matched the interests of the study participants, they were not always able to excite and motivate the pupils.

Secondly, technological shortcomings cause the interaction between user and virtual environment to be very limited. Commonly, the interaction revolves around explicit instructions from the users with the help of menus and buttons [Aylett et al., 2014; Sapouna et al., 2010]. While there are approaches for more "natural" interaction techniques, these are usually limited to scripted reactions to specific gestures [Damian et al., 2015a; Kistler et al., 2012]. Moreover, current speech recognition and natural language processing techniques are still far away from allowing a virtual agent to have natural conversations with the user in realistic scenarios. Thus, most systems simply perform voice activity detection to simulate turn taking, and completely ignore the content of the user's speech [Damian et al., 2015a; Gebhard et al., 2014; Pan et al., 2012]. These shortcomings drastically limit the credibility of the simulation. Moreover, when left alone with the system, the users might find ways to "break" the training exercise by taking advantage of these limitations.

Training systems also generally involve a rigid and time-consuming learning process, being designed around the concept that the learner goes through multiple repetitions to achieve results. Thus, the social skills training requires a significant time investment on the part of the user. Furthermore, the systems are usually stationary, requiring the presence of the user at a specific location (e.g. classroom). The reason for this is that they rely on complex sensing and processing equipment as well as controlled environments to run the simulation.

To sum up, although TARDIS and other technology-enhanced training systems are more accessible and scalable than coaching sessions with real human coaches, they still have a high barrier of entry. This dissertation introduces the concept of social augmentation as a more flexible and powerful form of social skills training. Social augmentation systems eliminate the need for setting up computationally expensive simulations and taking part in rigid a-priori training sessions. Instead, by relying on modern wearable devices, they allow users to boost their social skills during real social interactions anywhere and anytime. For instance, a social augmentation system can help a person better control their use of voice (e.g. loudness, speech rate) while speaking in public.

3.3 Going Mobile, Going Live

Computer-enhanced training systems represent the starting point for the work done in this dissertation. In fact, the concept of social augmentation actually originated from the idea of "adapting" the TARDIS system so that it can be used during real job interviews. Yet, social skills training systems are explicitly designed as primary tasks that a user carries out during dedicated training sessions in controlled environments. Transforming such a system to enable its use during actual social interactions involves overcoming several challenges.

The most obvious challenge involves the use of technology. In TARDIS, we relied heavily on the Microsoft Kinect to sense the user's behaviour and a Windows PC to process it and drive the simulation. Yet, the physical properties of these devices and their reliance on non-trivial setup processes make them impracticable to be used during spontaneous social interactions. Thus, a social augmentation system should be able to run on mobile devices and make use of unobtrusive wearable sensors.

Secondly, standard training practices involve giving the user feedback after the interaction during dedicated debriefing sessions. These debriefing phases usually present the behaviour analysis to the user with the help of a specialized graphical user interface (GUI). An example of such a GUI is illustrated in Figure 3.3. However, while such GUIs are appropriate for computer-based use, their complexity make them impossible to be used while participating in social interactions. Thus, another challenge is transmitting information about the behaviour analysis to the user, which can be processed in parallel to the social interaction without degrading the user's performance.

Finally, social interactions are governed by a set of unwritten rules to which all participants adhere without even knowing. The addition of new elements (e.g. the social augmentation system) to the interaction may be seen as interferences, disrupting the overall experience. To avoid this, the design of social augmentation systems needs to be respectful of the social etiquette. For example, to protect the privacy of the bystanders, which may be oblivious to the existence of the system, the augmentation should only use sensors that record and analyse the behaviour of the user.

3.4 Summary

To illustrate the state-of-the-art in terms of social skills training, this chapter provided the reader with a concrete example of a computer-enhanced training system. Starting from this example, the limitations of current training approaches are discussed. The social augmentation concept is then introduced as an alternative solution, one which overcomes the shortcomings of virtual simulations by allowing the training to happen during real social interactions.

The chapter ends with a discussion on how computer-enhanced training systems can be adapted to perform social augmentation, and the challenges they need to overcome to accomplish this. A more detailed discussion on the design and requirements of social augmentation is provided in Chapter 4.

II Concept

4. Augmenting Social Interactions

The overall incentive for augmenting social interactions is to help users participating in social interactions achieve a more advantageous outcome. For example, social augmentation could aid users deliver better and more compelling speeches by assisting them with their voice modulation or gesture usage. Job seekers with social dysfunctions could use social augmentation to land a better impression during job interviews. A social augmentation system could also be used to help persons who suffer from various disabilities, such as autism or Parkinson's disease better cope with social situations. Starting from these use cases, two concrete goals can be formulated.

1. *Generating awareness of one's own body.* As argued in the first chapter, it is often the case that one's perception of their own behaviour is not in line with how it is perceived by others. This is especially true during stressful situations such as speaking in public or job interviews, where stress and emotions can get in the way of our ability to control our gestures, postures or even speech. Certain disabilities are also associated with social dysfunctions, e.g. persons on the autistic spectrum often experience atypical prosody difficulties [Kanner, 1968] or people with Parkinson's have problems regulating their vocal loudness [Ramig et al., 2001]. An increased awareness of one's own behaviour would facilitate the detection and recognition of such unwanted behaviours, and thus represents the first step towards correcting them.
2. *Improving the quality of one's own behaviour.* Besides generating awareness, the social augmentation should also guide the user towards a behavioural state which benefits them more in the social interaction. For instance, during a job interview, the social augmentation should attempt to change the behaviour of the user with the aim of maximizing their chance at employment. To achieve this, the information delivered by the augmentation to the user should be sufficient to elicit a behavioural change. Yet, the augmentation must also be mindful of the limited and fragile nature of human attention and how too much information (or badly delivered information) can negatively impact the overall quality of the user's behaviour.

The remainder of this chapter will present the general concept of social augmentation. More specifically, in the following section, the concise requirements for augmenting social interactions are presented and discussed. Following this, the behavioural feedback loop is introduced as the driving mechanism behind social augmentation. The chapter then looks at how social augmentation can be customized to fit different scenarios and user types. Finally, the chapter provides a literature survey of related behaviour training systems while looking at both traditional training approaches and existing social augmentation systems.

4.1 Requirements

From the goals postulated in the beginning of the chapter, six concise requirements for a social augmentation system are devised. The first requirement relates to the augmentation's ability to deliver information to the user:

> **Requirement 1** The information delivered by the social augmentation can be correctly perceived and processed by the user.

It is important to note that the first requirement implies that the user is able to both perceive and decode all delivered information. For instance, if an audio-based augmentation uses pitch to encode information, it must take into account that not all users are equally proficient at telling different sound frequencies apart. Similarly, colour-based encoding can also be problematic since a sizeable part of the population suffers from colour blindness.

From a psychological point of view (c.f. Section 2.2), the augmentation represents a secondary task for the user whereas the social interaction is the primary task. Considering this, the first requirement can be reformulated as follows: The social augmentation is able to momentarily draw enough attention from the primary task to allow information from a social augmentation task to be perceived and processed.

However, it is crucial that the augmentation does not draw too much attention or else it would distract the user and disrupt the social interaction. According to distributive attention models [Navon and Gopher, 1979; Wickens, 2002], tasks can be carried out in parallel without quality degradation as long as enough processing resources are available. Thus, in order to reduce the amount of distraction, the social augmentation needs to be economical with its demand of resources. To this end, the second requirement is formulated as follows:

> **Requirement 2** The social augmentation has a minimal impact on the attention level dedicated to the primary task.

Yet, in order for the social augmentation to actually be able to guide the user to a more desirable behavioural state, it must be able to generally elicit a change in behaviour. Thus, the provided information must be understandable, sufficiently detailed as well as relevant to the user and the moment in which it is delivered. Informing a user who holds a presentation that they talked too loud five minutes ago is not only irrelevant but might also confuse the user. This leads us to the third requirement:

> **Requirement 3** The provided information is appropriate for facilitating the intended change in behaviour.

So far the social augmentation can trigger a change in the user's behaviour without

disturbing them too much. What is still missing is the relation to the social interaction mentioned in the beginnings of this chapter. More specifically, it is important that the augmentation does not just trigger any change in behaviour, but one that contributes to the goals of the user in the interaction. For example, in a job interview, the augmentation should help the user make a better impression, and thus increase their chances of employment. Considering this, the fourth requirement states:

Requirement 4 The intended behavioural change benefits the user in the social interaction.

The social aspect of the augmentation means that the physical form and aspect of the system is critical. For instance, a functionally perfect augmentation system will be rendered useless if the user has to wear a large "Darth Vader" helmet, as its presence alone would break social conventions and render any positive influence on the user's behaviour meaningless. It is particularly critical that the augmentation system does not hinder verbal or nonverbal communication within the social interaction. For example, the use of head-mounted displays (HMD) might prevent the perception of the user's gaze signals, interfering with one critical communication channel. Moreover, the augmentation must be mindful of its impact not only on the user, but also on the persons the user is interacting with. Thus, it is crucial that the augmentation system blends in as much as possible:

Requirement 5 The augmentation does not disrupt the social interaction or its participants.

Finally, privacy and transparency concerns also need to be addressed. Throughout history, many technologies found themselves at the receiving end of hate and anger over privacy concerns. When the first truly mobile camera (the Kodak camera) appeared in 1888, it was met with heavy criticism. David Lindsay's article on PBS.org[1] notes:

> *The appearance of Eastman's cameras was so sudden and so pervasive that the reaction in some quarters was fear. A figure called the "camera fiend" began to appear at beach resorts, prowling the premises until he could catch female bathers unawares. One resort felt the trend so heavily that it posted a notice: "PEOPLE ARE FORBIDDEN TO USE THEIR KODAKS ON THE BEACH." Other locations were no safer. For a time, Kodak cameras were banned from the Washington Monument. The "Hartford Courant" sounded the alarm as well, declaring that "the sedate citizen can't indulge in any hilariousness without the risk of being caught in the act and having his photograph passed around among his Sunday School children."*

Handling such privacy concerns is a delicate subject. One way to look at the issue is that the user pays for a service by sacrificing some of their privacy. To this end, as long as the transaction is fair, the user should be pleased. According to Hong [2013], the problem appears when the transaction is not fair, i.e. the receiving value does not match the value of the "lost" privacy. An especially critical case is when there is no benefit at all, as was the case for the Google Glass where "many people could be surreptitiously monitored by users of Google Glass at any time, and do not perceive any kind of value in return" [Hong, 2013]. Considering these issues, the final requirement states:

[1] http://www.pbs.org/wgbh/amex/eastman/peopleevents/pande13.html

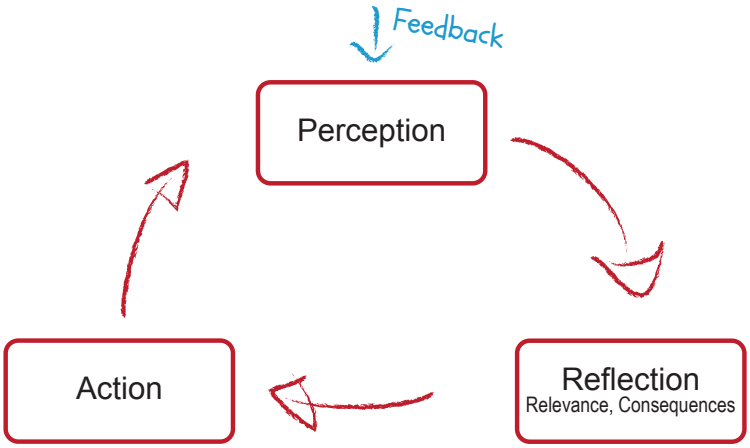

Figure 4.1: The three main phases of a feedback loop from the user's point of view.

Requirement 6 The social augmentation respects the privacy of both the user and the bystanders.

4.2 Behavioural Feedback Loop

Feedback loops present themselves as a natural choice for the driver behind social augmentation. In the simplest of terms, a feedback loop occurs when the output of a system is repeatedly and continuously fed back to the system as input, thus forming a closed loop. From the point of view of social augmentation, feedback loops are particularly interesting due to their self-regulating nature. Thus, the goal of generating self-awareness presented in beginning of this chapter can be directly translated to a feedback loop structure. The user's behaviour (output) is recorded and fed back to the user (input) continuously, generating awareness of one's own behaviour. Now, through intelligent and goal-oriented manipulation of the feedback loop, the behaviour of the user can be steered towards a more beneficial state for the social interaction. This thesis refers to this specialization of the general feedback loop as a behavioural feedback loop.

The feedback loop lies at the core of various psychological models, such as observational learning [Shettleworth, 2009], operand conditioning [Skinner, 1938] or social cognitive theory [Bandura, 1986]. From a user's point of view, a feedback loop involves three main phases: perception, reflection and action. In the perception phase, the user acquires information from an internal or external source. This information is then processed by the user to evaluate its relevance and the consequences of acting upon it. Finally, if the information has been deemed relevant and the consequences profitable, an action is executed in effort to shift to a more advantageous state within the environment. This progression is illustrated in Figure 4.1

One very successful example of such automated feedback loops are dynamic traffic speed displays. These technology-enhanced traffic signs include a radar and a display to give drivers realtime feedback on their speed. They were first used in Garden Grove, California out of frustration over the poor success rate of conventional methods to change driving behaviour near schools. To the surprise of everyone, despite their simplicity and redundant nature (they provide the drivers with information they already know), average speeds decreased considerably. In this example, in the perception phase, the driver sees the display and extracts

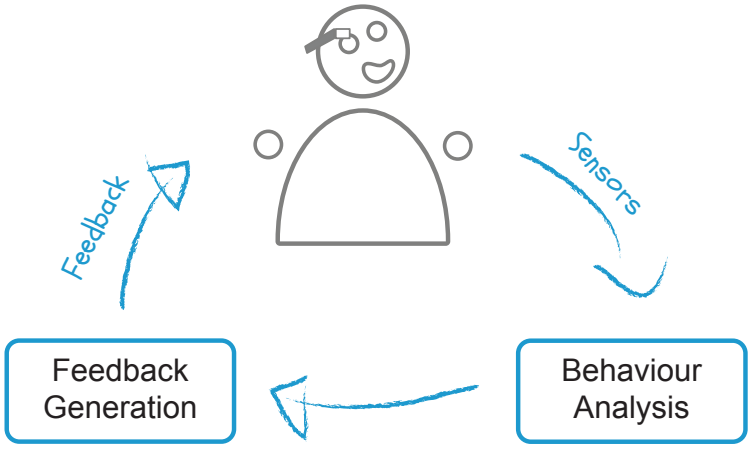

Figure 4.2: The behavioural feedback loop featuring its two main components: behaviour analysis and feedback generation.

the presented information. The driver then analyses the information and recognizes it refers to their behaviour, therefore deeming it relevant. Finally, the consequences of the information (e.g. possible accident or speeding ticket) are analysed and an action is executed (reduces the speed of the vehicle) to change to a more desirable state (safety).

This dissertation looks at behavioural feedback loops relative to their power to change social behaviour by increasing the user's awareness of their own behaviour, providing information on the quality of the current behaviour and making clear what actions are needed to shift to a more desirable behavioural state. To achieve these effects, a two-step pipeline (see Figure 4.2) is proposed: The user behaviour is first analysed in realtime and then, based on its quality, feedback is automatically generated and delivered to the user.

4.2.1 Behaviour Analysis

The first step of a behavioural feedback loop is the analysis of user signals. This step is responsible for extracting, processing and feeding behavioural data into the loop, allowing custom feedback to be generated and delivered to the user. What behaviour is analysed depends on the user and the scenario. Possible options include speech rate, loudness, gestures and postures.

Two social augmentation requirements (see Section 4.1) are directly related to the behaviour analysis component of the loop: R3 and R4. Both requirement imply that the feedback needs to be relevant to the user, the social interaction and the moment in order to facilitate a meaningful behavioural change. Thus, the behaviour analysis component needs to be adaptable to cater to different users and scenarios as well as robust towards environmental factors. For example, a gesture performed while sitting in front of a computer should be just as well processed and analysed when the user is walking down the street. Moreover, to guarantee that the feedback is relevant to the moment, the behaviour analysis needs to be performed continuously in realtime and with minimal latency.

The fifth and sixth requirements also relate to the design of the behaviour analysis. More specifically, the analysis process needs to be mobile and not constrict the behaviour and movement of the user (R5). Thus, the sensing and processing units need to be small, lightweight and inconspicuous. Furthermore, the behaviour analysis needs to respect the

privacy of the user and the interlocutors (R6). It should only record and analyse the behaviour of the user, since only the user receives a direct benefit from the system. Long term data storing should also be avoided unless explicitly requested by the user. If data storing is indeed needed (or requested), only processed and anonymous data should be stored.

All these requirements can be fulfilled by delegating the behaviour analysis to a mobile social signal processing pipeline. Chapter 5 will provide a more in-depth discussion on the challenges and possibilities of mobile social signal processing. A concrete implementation of a mobile social signal processing framework is presented in Chapter 7.

4.2.2 Feedback

The second phase of a behavioural feedback loop handles the generation and delivery of feedback to the user. The feedback is generated using information provided by the behaviour analysis and by taking into account user and scenario particularities. It is then delivered using one or multiple channels, e.g. visual, auditory or tactile. For this, various output devices such as head mounted displays, headphones or vibro-actuators can be used.

The concept of giving the user information regarding a specific event while prompting them to alter their behaviour, resembles the classical notification mechanism of modern computer systems. Similarly to notification systems, the feedback is potentially interrupting to the main task. Users need to sacrifice attention from their primary task to receive additional information from feedback. Thus, intelligent feedback design is critical for the fulfilment of all social augmentation design requirements (see Section 4.1).

McCrickard [McCrickard et al., 2003] identified three design principles for notification systems: interruption, reaction and comprehension. The first principle is in line with our second requirement and attempts to diminish the amount of attention the users need to sacrifice from their primary task. This attention covers visual attention (focus) but also mental attention and the distribution of processing resources (see Section 2.2). The second and third principles state that a notification should be perceivable and comprehensible by the user, and able to elicit a reaction. These two principles can be mapped to R1 and R3. More precisely, as postulated in R1, the feedback needs to be understandable at a glance and distinctive enough to bypass any perceptual bottleneck (see Section 2.2.2). The feedback also needs to provide the user with clear and sufficient information to facilitate a change in behaviour (R3). R4 and R5 extend McCrickard's principles to also include guidelines regarding the goal of the feedback and its impact on the user and bystanders. That is, feedback should follow the goal of improving the quality of the user's behaviour from the point of view of the interaction (R4). Moreover, similarly to the behaviour analysis, the choice in hardware is important to ensure a minimal impact on the situation and the user (R5). Large and heavy head-mounted displays (HMD) or headphones should be avoided, and wireless devices are preferred. Finally, to appease the sixth requirement and respect the privacy of the user, the feedback should only be perceivable by the user. For example, feedback should be delivered using head mounted displays or headphones instead of monitors or loudspeakers.

A feedback event is characterized by various factors: modality (Which channel is the feedback sent over?), duration (How long is the user exposed to the feedback?), prominence (How "flashy" is the feedback?), scope (Does the feedback instruct the user what do to or does it only present the current situation?) and level of detail (How detailed is the feedback?). All these factors contribute to the impact of the feedback on the user and are thus critical to the effectiveness of the whole loop. Chapter 6 discusses the feedback factors and their role

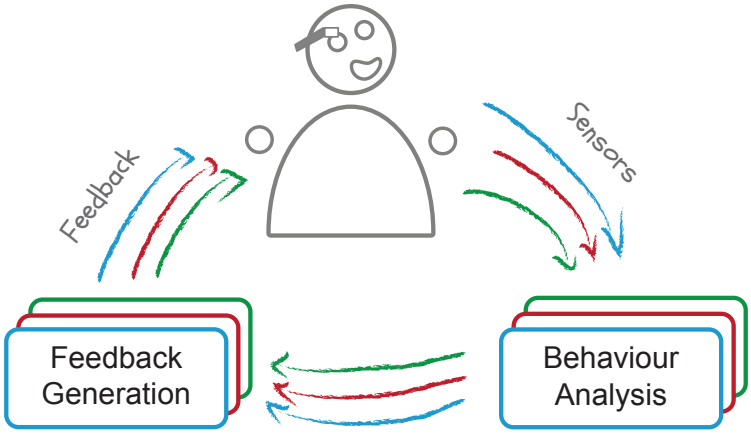

Figure 4.3: Multiple behavioural feedback loops.

within social augmentation more in detail.

4.2.3 Multiple Loops

So far in this chapter, the behavioural feedback loop has been described as a continuous loop where one signal starts at the user, goes through the behaviour analysis and feedback generation stage and ends back at the user. While never explicitly stated, it was somewhat implied that there is always one active loop at a time. In this section, the case when this is not true will be discussed: What happens if there are multiple loops running in parallel? and what are the benefits and drawbacks of multi-loop setups? Figure 4.3 illustrates the idea.

At its core, a multi-loop setup is defined as two or more behavioural feedback loops which run in parallel and involve the same user but different behaviour analysis and feedback generation stages. For example, a social augmentation system that analyses and provides feedback on both gesture and speech quality. Here, two loops, both involving the same user, are required. The first takes data from the body tracking sensors to analyse various features of the user's body and then provides feedback to the user in accordance to the results of the analysis. The second loop polls data from the microphone and performs audio analysis on it. Based on this audio analysis, feedback is generated and delivered to the user in parallel to the gesture feedback of the first loop.

The main advantage of using multiple loops is providing the user with more information on their own behaviour, and thus achieving a higher behaviour improvement potential. More precisely, augmentation strategies can be devised across all loops at the same time, allowing the individual loops to work complementary to one another. The most obvious type of complementarity is augmentation broadness. Here, multiple loops work together to provide a broader coverage of social behaviours (e.g. speech, gestures, facial expressions) and allow the user to improve multiple behaviours at the same time. Another strategy involves the use of multiple behaviour feedback loops to cover different temporal planes. For instance, one loop can provide moment-by-moment low-level instructions whereas another loop may give the user a more global view of their performance.

However, similarly to how a secondary task (i.e. the feedback perception) can impact the execution of the primary task and vice-versa (see Section 2.2.1), in a multi-loop scenario, each loop acts as a separate task and may impact the execution of the others. To explain

this effect, the term interference is borrowed from Kahneman [1973]. According to him, an interference between two tasks occurs when both attempt to use the same cognitive resource. Whereas Kahneman only looked at cognitive interferences, the following paragraphs will use the term more broadly to discuss all types of interferences which can occur when dealing with multiple feedback loops:

- *Sensor Interferences*. These are usually physical interferences caused by the shape and size of the sensing hardware. For example, two head mounted sensors, such as an eye tracker and a microphone, are very likely to interfere with each other because of the physical characteristics of each sensor and how it is attached to the user's head. Function-related interferences are also possible. For example, inertial measurement units (IMUs) are susceptible to magnetic and electric disturbances. Thus, such sensors should not be placed in the vicinity of hardware which can cause such disturbances. Finally, the most common type of sensor interferences are related to their physiological and psychological effect on the user. On the physiological side, the weight and size of one sensor can constrict the movement of the user which in turn can affect what other sensors are measuring. Similarly, the mere presence of the sensors might impact the state of mind of the user, and thus their behaviour (see also discussion in Section 5.2).
- *Incompatible Feedback Designs*. The feedback generation phase is also susceptible to interferences caused by incompatible feedback designs. For example, while it is possible for multiple loops to have different prominence levels, such setups would most likely be detrimental to the goals of each individual loop. A highly prominent (e.g. loud) auditory feedback will undermine any attempt of a secondary loop to be subtle and undisruptive.
- *Output Interferences* occur when two loops opt to use the same output device (e.g. Google Glass) or two incompatible output devices. Using the same output device needs careful synchronization of feedback events, both in terms of timing and content to avoid having one event disrupt the perception of the others. In case of using multiple output devices, similarly to sensor interferences, the shape and size of the hardware can cause physical, functional, physiological or psychological interferences. Moreover, interferences between sensor and output devices are also possible. For example, a head-mounted display may interfere with a head-mounted eye tracker.
- *Cognitive Interferences*. In this context, each feedback event represents a different task the user's cognitive system needs to process. If the perception of, or reaction to two or more feedback events overlap, distributed attention mechanisms are activated. This can impact the performance for handling these events, e.g. in the case of resource insufficiency. Furthermore, the individual feedback processing tasks can also impact the performance of the primary task, i.e. the social interaction.

Regardless of the type of interference, it will most likely have an effect on the user and the augmentation process. It is to be expected that whenever an interference occurs between two loops, the performance of each loop as well as that of the primary task decreases. Thus, a careful design of multi-loop setups is paramount.

4.3 Adapting to User and Context

Humans are complex beings operating in a complex world. A visual message might be correctly perceived and decoded by one person but completely ignored by another one, simply

because for one of the persons the stimulus resembled the poached egg she had for breakfast on the first day of her honeymoon. A beep might be flawlessly perceived at the beginning of a social interaction but missed ten minutes later because of an increase in background noise. Thus, the feedback (as well as the whole behavioural feedback loop) needs to be able to adapt to the user, the context and the scenario. This is similar to the problem described by Arroyo et al. [2002]. While studying how users respond to different types of stimuli, they found that "more notable than the differences between modalities was the differences between people" and that "subjects' sensitiveness depended on their previous life exposure to modalities."

To tackle this issue, a three forked adaptation approach is proposed. First, based on the user activity, the social augmentation is automatically turned on or off. Second, once the augmentation is active, realtime adaptation techniques are employed to allow the behavioural feedback loop to continuously adjust to the user's behaviour during the social interaction. Finally, low level mechanisms manage the timing of the individual feedback events to make sure they are delivered at opportune moments.

4.3.1 Augmentation Activation

The vision behind the proposed augmentation concept is that of a system which accompanies users throughout the day, helping them deal with new social encounters and providing tips on how to handle situations they are not proficient at. Yet, to make such a system viable for day-to-day use, it must be able to automatically engage and disengage depending on whether the user is currently participating in a social interaction or not. This ability not only reduces the system's energy footprint, but also minimizes the probability of behaviour classification errors or inaccurate feedback events, and helps alleviate privacy concerns.

However, automatically detecting whether the user is participating in a social interaction is a difficult task. Multiple approaches have been proposed in related literature. The simplest approach relies on measuring the interpersonal distance between pairs of users. If this distance drops below a certain threshold, the presence of a social interaction is likely. A more in-depth discussion on this technique is provided in Section 5.3.5. However, small interpersonal distance does not guarantee the presence of a social interaction. Just think of yourself riding a bus at rush hour and you will quickly understand why this is the case. To tackle this issue, Matic et al. [2012] also use body orientation and speech activity in addition to interpersonal distance to perform social interaction detection.

An alternative approach is to use front-facing cameras (such as the one found on head-mounted displays) for performing real-time face tracking on the video stream. This would allow the system to detect the presence of mutual gaze which is a prime indicator of an ongoing interaction. Yet, this solution might raise privacy-related concerns and goes against the social augmentation's principle of only analysing the user's own behaviour.

Once evidence of a social interaction is found, the augmentation is automatically turned on. Yet, it is likely that different interactions require different types of augmentation. For example, when speaking in public the user should be encouraged to be energetic in an effort to stimulate the audience. This is not the case for face-to-face interactions (e.g. during a job interview), where energetic behaviour is generally considered inappropriate. Thus, it would be of benefit if not only the presence of a social interaction could be detected, but also its type. This way, the system would be able to initiate an augmentation which is appropriate for the interaction. Unfortunately, a literature survey revealed no concrete examples for automatic social interaction classification techniques which go beyond simple binary detection (see also

discussion in Section 10.2.3).

4.3.2 Online Adaptation of the Feedback Strategy

Once the augmentation is active, two different adaptation mechanisms are used to further enhance the effectiveness of the behavioural feedback loop. First, a feedback validation stage enables the system to monitor the impact of the feedback on the user. After each feedback event, the validation stage measures how the user reacted to the feedback and adjusts the next feedback event if no meaningful response from the user can be detected. For instance, if the validation of a feedback event fails, the prominence of the next event is increased (e.g. the event becomes louder, uses brighter colours or stronger vibrations).

This mechanism has the advantage of permitting the fuzzification of the edges of the behaviour classification. Small "errors" in behaviour result in low-intensity, subtle feedback, whereas persistence in an inappropriate behaviour would gradually cause ever stronger feedback until the user adjusts. A concrete example for feedback validation has been proposed in [Damian and André, 2016]. Here, the intensity of the vibrotactile feedback is increased every 10 seconds if the user does not adapt their behaviour.

The second proposed online feedback adaptation mechanism is modality progression. More precisely, at the beginning, each feedback is unimodal. Then, depending on how well (or, better said, badly) the user responds to the feedback (as measured by a validation stage), additional modalities can be used to increase the likelihood of triggering a behavioural change. Once the user reacts, the system switches back to an unimodal strategy to leverage feedback effectiveness against perceptual overload. For instance, initially, visual feedback can be used to inform the user that they are speaking too fast. If the user does not correct their behaviour, auditory feedback can be triggered in addition to the visual one until the user corrects their behaviour. How exactly multimodal feedback can be used within a behavioural feedback loop is discussed in Section 6.1.7. Section 7.5.4 then provides a concrete implementation of modality progression.

4.3.3 Timing Management

Finally, during an active augmentation, timing management controls when a feedback event is actually delivered to the user. Studies have found that in a multitasking scenario, the timing of an interruption by a secondary task has a large influence on the disruption of the primary task. In our case, starting from the assumption that the user will not always be able to perceive and interpret the feedback completely parallel to the primary task, the "when" of the feedback delivery is crucial in limiting (and perhaps even preventing) interruption-induced performance degradation of the primary task. More specifically, timing management can affect the cost and the frequency of interruptions. Whereas frequency is a fairly straightforward concept measuring the amount of interruptions per time frame, the cost of an interruption is a measure for the decrease in performance of the primary task caused by the interruption [Cutrell et al., 2001; Gillie and Broadbent, 1989]. In their work, Horvitz and Apacible [2003] associate this cost with the user's willingness to pay real money for avoiding the interruption.

One approach towards managing the timing of feedback is to reduce the overall amount of feedback events, and thus diminish the frequency of possible interruptions. The social augmentation application described in Chapter 8 follows this approach. More specifically, the signal processing pipeline handling the behaviour analysis uses large window filtering to avoid the triggering of feedback events by high frequency spikes (e.g. noise or sudden changes in

behaviour). Moreover, the classification of the behaviour uses a fuzzy thresholding approach which favours a slower but more stable classification. In the group discussion augmentation system presented in Chapter 9, we used a simpler approach and limited the overall amount of feedback events the system sends to the user. Another common technique is to impose a timeout window following each feedback during which no new feedback is triggered [Boyd et al., 2016; Schneider et al., 2015; Tanveer et al., 2015].

More complex timing management strategies are those which not only reduce the frequency of possible interruptions but also actively attempt to diminish their cost. Horvitz and Apacible [2003] propose using a dynamic Bayesian network for automatically predicting the cost of an interruption during typical office activities. The Bayesian network receives as input various user activity events including high-level (e.g. switching between applications, closing a file, calendar events) and low-level events (e.g. user gaze direction, voice activity). Similarly, Poppinga et al. [2014] and Okoshi et al. [2017] also propose systems for classifying interruption cost. However, unlike the work by Horvitz, they attempt to determine opportune moments for delivering notification messages (e.g. news headlines) on smartphones. For this they rely on smartphone sensors such as phone orientation, phone location (indoor or outdoor), time of day, activity type (e.g. walking, standing) or screen status (on or off). Generally, the concept of delivering feedback only at moments where the cost of interruption is minimal is promising. However, since previous research is highly scenario-specific, more work is needed to identify pertinent behavioural events and efficient classification methods for the social augmentation context (see discussion in Section 10.2.4).

A different approach has been presented by Bailey and Iqbal [2008]. According to them, interruptions are less disruptive if perceived at task boundaries where cognitive load reaches a local minimum. Thus, one possibility is to deliver feedback only at such task boundaries. However, what exactly these boundaries are is difficult to say and heavily dependent on the scenario. One type of task boundaries is transitions between goal-oriented processes such as saving a document [Iqbal and Bailey, 2005] or typing in a search query [Czerwinski et al., 2000]. In a social augmentation scenario, task boundaries can be of two types. First, low-level task boundaries are pauses in continuous speech. On the other hand, high-level task boundaries are moments of transitions between phases (e.g. discourse segments, slide switching during a presentation, proceeding to a new question in an interview). Both types of task boundaries have advantages and disadvantages. Generally speaking, high-level boundaries are more likely to correlate to moments of low cognitive demand [Iqbal and Bailey, 2005; Miyata and Norman, 1986]. In computer-assisted or computer-mediated scenarios, external systems can send interaction context information to the feedback timing manager, which can be used to infer high-level boundaries. For example, in the job interview training application presented in Section 3.1, the scenario manager can send information on the current interview phase to the feedback manager. In this case, a switch from one phase to another can be interpreted as a high-level boundary. However, information about high-level boundaries is not always available especially when working in an in-situ social environment. Thus, in such situations, low-level boundaries can act as an alternative since they are easier to detect.

Simpler approaches for managing interruption cost have also been proposed. Cutrell et al. [2001] found that feedback received at the beginning of a task was more disruptive than feedback received later. Thus, "it may be less disruptive in some situations to delay the transmission of notifications." In the social augmentation scenario, this effect can be used by simply avoiding to provide feedback during the initial phases of an interaction (e.g. in the

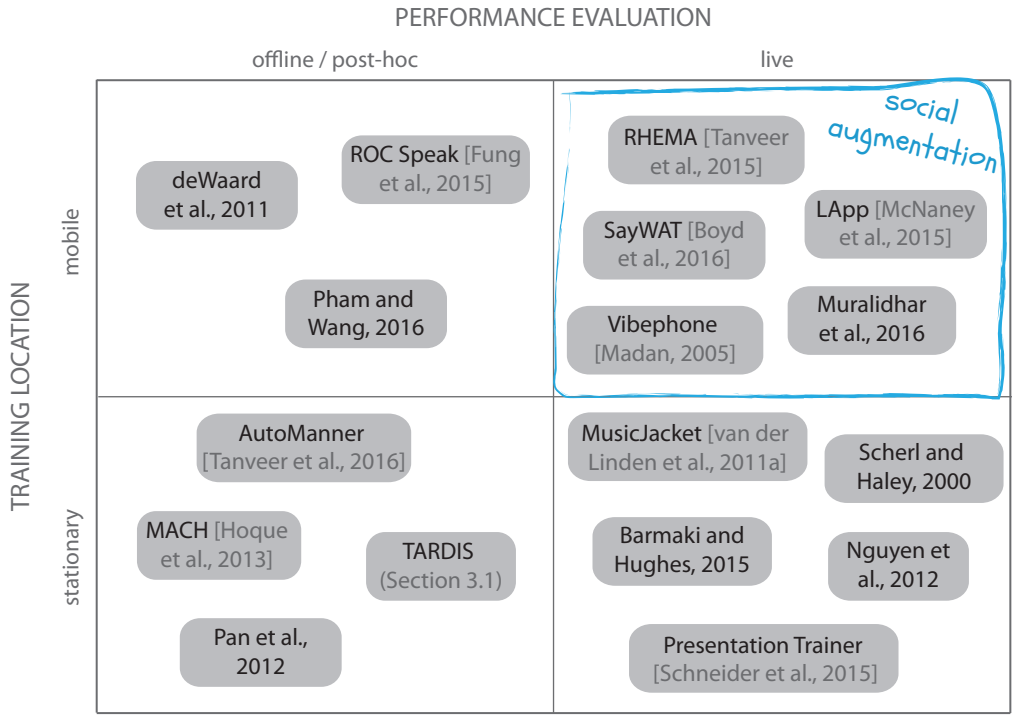

Figure 4.4: Social augmentation in relation to current training practices and related work.

beginning of a public speech).

Another possibility of managing the timing of the feedback is by taking into account the temporal "blind spots" of the human perception systems. Delivering feedback during these perceptual pauses may diminish the intensity of (or completely avoid) the interruption. This approach has been taken by Vidal et al. [2014]. The authors use an eye tracker to determine where the user is looking and when they blink. Using this system, they propose two concepts for reducing interruptions caused by updating visual elements. First, they propose to update the colour of visual elements on the display only when the user is not looking at the display, making use of the peripheral vision's poor colour perception [Wooten and Wald, 1973]. Second, they propose to perform larger changes to the visual display while the user is blinking and thus rely on the change blindness effect [Davies and Beeharee, 2012]. Thus, unlike the approaches discussed before, Vidal does not attempt to reduce the cost or frequency of interruptions, but to eliminate involuntary attentional capture altogether. This subtle difference has large implications in the context of social augmentation. More specifically, without involuntary attentional capture, the user will only notice the feedback when they purposefully direct their focus towards the display. In a highly stressful situation, this might not happen very often (if at all), causing the user the potentially miss the feedback, violating R1.

4.4 Related Work

In order to give the reader a better overview of how social augmentation is positioned relative to current training approaches, this section will provide a classification scheme for training systems, one which takes into account where the training is happening and how

the performance of the learner is evaluated. This should give the reader a clearer picture of the role of social augmentation in today's "training ecosystem" and how it compares to other systems. Figure 4.4 illustrates this classification scheme and places the proposed social augmentation concept in relation to known training practices and related work.

The first dimension, performance evaluation, classifies training methods according to when the performance evaluation occurs, and thus when the learners receive feedback on their behaviour. Two classes emerge: offline and live. Offline performance evaluation represents the classical two-phase training approach where learners first absorb knowledge and then receive feedback on how well they were able to apply it (e.g. following a written or oral test). The offline performance evaluation has the benefit of not having any time constraints, allowing a single teacher to evaluate multiple students thoroughly, or a computer program to perform computationally demanding behavioural analysis. In contrast, a live performance evaluation means that the two training phases overlap. Thus, the learner's behaviour is continuously analysed and feedback is provided immediately. This drastically reduces the duration of the learning cycles since the learner can apply the feedback directly and see the results instantly.

The training location dimension classifies training approach into two categories, stationary and mobile. Stationary training approaches are location-bound and require the physical presence of the learner. This allows the training to happen in a controlled environment, where the impact of environmental factors can be minimized. Yet, this controlled nature reduces the realism of the training, whereas the fixed location restricts the learner's exposure to the training material and impacts the overall user experience. On the other hand, mobile training does not restrict the mobility of the users, allowing them to train at a time and place of their own choosing. Mobile training also enables in-situ training, i.e. training in the same environment where the practising happens.

In the upper right corner of the classification scheme, at the crossroads between mobile training and live feedback, lays the social augmentation concept. It combines the flexibility and accessibility of mobile training with the convenience and swiftness of live performance evaluation. This allows the training to be performed during the actual interactions one wishes to become better at. The remainder of this section will go into detail on each training type and provide examples from related literature.

4.4.1 Offline and Stationary Training

Commonly, computer-enhanced training systems rely on desktop PCs and offline performance evaluation for helping users improve their social skills. They strongly resemble the classical two-phase learning paradigm according to which the user first takes part in learning exercises, and then receives feedback on what they did good or bad.

A typical example of such a system is TARDIS (see Section 3.1). It makes use of simulation techniques to place the user in a virtual job interview. After the interview, the user receives feedback on their use of behaviour and how it might have influenced the outcome of the interaction. MACH [Hoque et al., 2013] is a similar system that also places the user in a simulated job interview scenario. At the centre of the simulation is a virtual agent that is capable of asking questions and displaying nonverbal behaviour. After each interview, MACH provides the user with feedback on their use of language, facial expressions and head movements.

Another example has been proposed by Tanveer et al. [2016]. Their system, AutoManner, aims at helping users reduce the amount of unconscious body movement (referred to as

mannerisms) they perform while speaking in public. For this, the users are asked to practice their public speech in front of a Microsoft Kinect. The system then analyses the data for recurring patterns of body movement, and provides the user with feedback on the amount and distribution of mannerisms during their talk.

4.4.2 Live Performance Evaluation

Live feedback as a concept is not new and has already been successfully used as a coaching instrument in certain domains. Hypnosis therapist would observe and provide feedback to trainees during interventions with real patients by sitting next to the trainees. As hypnosis lost in popularity and confidential psychoanalytical interventions became more common, it was no longer appropriate for the expert therapist to sit in the same room with the trainee [Scherl and Haley, 2000]. To solve this issue, one-way "magic" mirrors would be used to allow the therapist to observe the trainee. Initially, to provide feedback to the trainee, the therapist had to ask the trainee out of the room or call on the telephone whenever they noticed an error. This was distracting to the intervention and revealed to the patient that something was amiss. This issue was partially solved with the arrival of the "bug in the ear" [Neukrug, 1991]. This small earphone was placed in the trainee's ear and allowed the expert to easily provide feedback and guidance. However, this solution was prone to distract the trainee, potentially causing "glazed eyes" and making the arrangement become akin to "a robot being instructed" [Scherl and Haley, 2000]. Another approach to this issue was presented by Scherl and Haley [2000]. They placed a computer monitor in the room of the intervention only viewable by the trainee and allowed the therapist to write messages on it. An evaluation revealed that using this setup, feedback could be provided effectively during conversations between trainee therapists and their patients as long as the directives from the human supervisor are clearly formulated and kept short.

The first automatic live feedback systems only started to appear well after the turn of the century, together with the miniaturization of sensing devices. MusicJacket [van der Linden et al., 2011a,b] used tactile feedback to train pupils to correctly play the violin. For this, it used an assembly of inertial sensors and vibrotactile actuators shaped in the form of a jacket. The system automatically analyses the position and the motion of the user's arms and triggers tactile feedback once it detects erroneous behaviour. MusicJacket scored well during user studies when compared with traditional teaching methods, enhancing the motor learning process.

Live feedback systems have also been developed for public speaking training. One of the first systems of this kind was developed by Nguyen et al. [2012]. They used a Microsoft Kinect to analyse the user's gestures and postures. Once a "bad" gesture or posture is detected, a visual feedback is provided on a monitor. A similar concept has been proposed by Schneider et al. [2015]. Their "Presentation Trainer" uses both visual and tactile feedback to alert the user to "presentation mistakes" such as bad postures, insufficient use of gestures, lack of speaking pauses and inexpedient loudness. The concept has been picked up by other researchers who extended it with multimodal behaviour analysis [Dermody and Sutherland, 2015, 2016] or applied it in different scenarios [Barmaki and Hughes, 2015]. Researchers also experimented with interactive virtual audiences for delivering social live feedback (i.e. nods and posture changes) to the user [Chollet et al., 2015]. They found that the virtual audience outperforms visual feedback shown on a screen.

The job interview training system described in Section 3.1 also incorporated live feedback

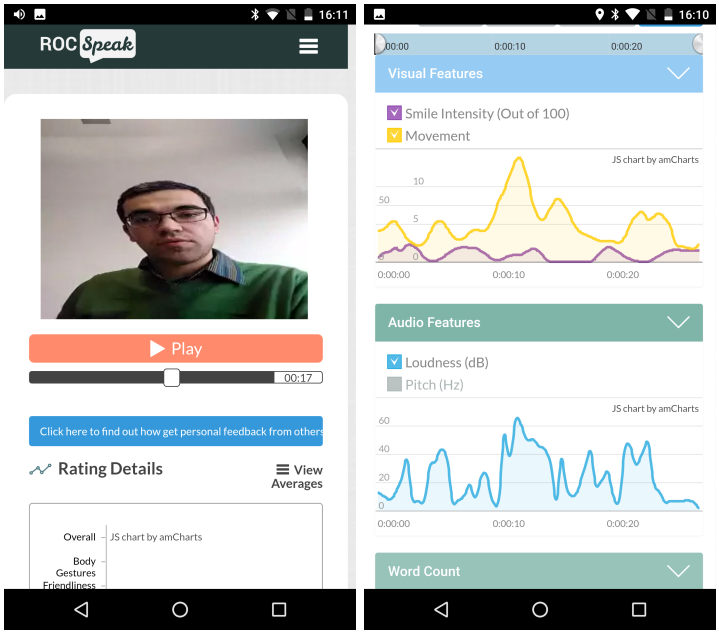

Figure 4.5: The ROC Speak web application for training public speaking skills.

elements. The system's GUI showed a series of icons which informed the user which behavioural tasks on the game cards have been completed. These visual elements increased the user's awareness of their own behaviour. Furthermore, the virtual character was also able to react to the user's behaviour by showing backchannelling animations such as gazing or nodding. These signal that the virtual character is actively participating in the conversation and enhance the realism of the simulation.

While, all of these systems are capable of realtime performance evaluation, they are still stationary since they require the learner's presence at a specific location and enforce a controlled environment. Furthermore, the feedback used in most cases is crude and does not take into consideration the state of the user's attention. The next section goes beyond this and discusses systems which can be used "in the wild."

4.4.3 Mobile Training

The second axis of the proposed classification scheme concerns itself with the training location. Namely, it distinguishes between training approaches which enforce fixed locations and controlled settings, and approaches which allow the learner to train on-the-go.

Most training approaches are stationary. They require the learner to be physically present at a specific location (e.g. classroom), minimizing environmental influences (e.g. sound insulation). Examples can be found in both traditional human-assisted (e.g. school, therapy) and computer-enhanced learning [Anderson et al., 2013; Schneider et al., 2015; van der Linden et al., 2011b]. Especially for computer-enhanced systems with live performance evaluation, a fixed location means that powerful computers can be used to perform complex behaviour analysis tasks. Furthermore, a controlled environment allows researchers to use sensitive sensing equipment since they are able to carefully calibrate it without worry from external factors.

With the advent of internet-connected smartphones, learners gained access to a seemingly infinite pool of knowledge right from the palm of their hands. Using such devices, various

learning experiences have been proposed which do not constrain the mobility of the user. These systems offer an unprecedented amount of flexibility, allowing users to learn wherever and whenever they want. Typical examples of mobile learning systems are mobile Massive Open Online Courses (MOOC) which can be used to acquire both social and professional skills. While traditionally MOOCs are accessed from stationary personal computers, recent efforts have looked at how to bring MOOCs to mobile platforms [Koutropoulos et al., 2011; Pham and Wang, 2016].

Another example of a system capable of performing mobile training is ROC Speak [Fung et al., 2015]. The system represents a more streamlined version of the classical computer-enhanced training approaches, which allows the user to participate in training sessions directly from their internet browser. This makes it usable from both personal computers and mobile devices. The aim of ROC Speak is to help users improve their public speaking skills. For this, the users need to first record themselves giving a speech using either a webcam on a computer, or the camera of a smartphone. The recording is then sent to a server for analysis. Once the analysis is finished, the user receives feedback on their use of language, facial expressions and body movement (see Figure 4.5). In addition to the automatic analysis, ROC speak also offers the option of contracting Mechanical Turk workers[2] for analysing the video and providing subjective feedback.

4.4.4 Social Augmentation

Finally, social augmentation lies at the crossroads between mobile training and live feedback. One of the first systems to attempt social augmentation was Madan and Pentland's Vibephone [Madan, 2005; Madan and Pentland, 2006]. It analysed the user's voice to extract various paralinguistic features such as speaking time, voice rate, and prosodic emphasis. The system then provided visual feedback on the mobile phone's display on the current status of these features as well as a prediction on the overall quality of the interaction (in this case, a romantic date). However, there are very few details available on the system and no evaluation of any form has been published.

A more modern (and elaborate) take on social augmentation has been proposed by Tanveer et al. [2015]. Using the Google Glass, they implemented the RHEMA system for assisting users while delivering a public speech. At the centre of their concept lies the idea that "the speaker needs to modulate his or her volume and vary speaking rate to retain the audience's attention." For this, they stream the audio signal from Google Glass's microphone to a server where it is analysed in terms of loudness and speaking rate. Based on the results of this analysis, they support the user in achieving a better modulation by displaying visual feedback on the Google Glass.

Another system which can be classified as social augmentation was proposed by Boyd et al. [2016]. Named SayWAT, their system targets adults with autism and provides support for improving their prosody during social interactions. Similarly to Tanveer et al. [2015], they analyse Google Glass's audio stream in terms of loudness and pitch variation. If one of these values crosses a predefined threshold, they trigger a visual feedback on the Google Glass. A study revealed that users were generally able to perceive the feedback without distraction and quantitative results suggest that the behaviour quality, especially the loudness, improved due

[2]Mechanical Turk is a service offered by amazon USA, which enables the easy recruitment of persons (called Turks or Turk workers) for short activities (called jobs), such as participating in online studies or analysing videos. The Turks receive financial compensation for each completed job.

to the system.

McNaney et al. [2015] proposed a similar system, but targeted people with Parkinson's (PwP). One common problem PwP face is that they "have an impaired perception of how loud they are speaking." Thus, the authors implemented the LApp system. It uses a Google Glass to continuously monitor the loudness of the user's voice, and give visual feedback whenever it drops below a predefined threshold. A user study showed that the system was well received and appeared to have a positive effect on the participants' ability to regulate their loudness.

Muralidhar et al. [2016] also use a Google Glass to provide sales apprentices with feedback on how much to talk during conversations with customers. For this, their system analyses how long the user speaks within a 20 second window, and provides textual visual or auditory feedback if the user should speak less or more. Subjective results from a small scale user study suggests the system was considered useful and not distracting.

A different type of system, which is related to social augmentation, is sensory substitution systems. They are targeted at persons with disabilities (e.g. visual impairment) and aim at substituting one sensory channel with another. The most common ones are sonification systems. These encode information sighted persons commonly perceive using the visual channel, into sounds. Thus, visually impaired persons can perceive visual information using their sense of hearing. In most cases, sonification systems target basic information types, such as colour [Banf and Blanz, 2013; Bologna et al., 2009; Dietz et al., 2016] or depth [Twardon et al., 2013], and aim to help the user navigate the environment or recognize objects. However, there are approaches [Dietz et al., 2016; Guizatdinova and Guo, 2003; Patil et al., 2012] which sonify social information and assist the visually impaired in their social interactions.

Overall, there are already some examples of systems which fall in the category of social augmentation. Most of these have been implemented concurrently to the writing of this thesis, and draw from the ideas and concepts introduced in Damian et al. [2014a] and Damian et al. [2015b]. However, a large portion of these systems suffer from inconsistencies in terms of design, implementation and evaluation. For example, Tanveer et al. [2015] report that users prefer textual instructional feedback over symbolic appraisive feedback. Yet, it is not clear whether this difference in preference was caused by the information representation (textual versus symbolic) or the feedback scope (instructions for change versus appraisal of current state). Furthermore, the study did not measure the impact of the feedback on a live audience. It is debatable whether a Mechanical Turk watching a video recording of a public speech is, for example, able to correctly ascertain the degree to which the speaker "retains the audience's attention." Similarly, Boyd et al. [2016] used two types of feedback, one symbolic (for loudness) and one textual (for pitch), yet the authors fail to explain the lack of consistency. The feedback is also appraisive in nature in a sense that it does not provide instructions on how to change the behaviour, it only displays the current behavioural state of the user. Although the authors initially argue that this makes the feedback clearer, easier to interpret and fosters self-efficacy, the results of their user study suggest otherwise, with participants experiencing "difficulty in taking action" and requesting "directives" and "directions." Nevertheless, these initial attempts at augmenting social interactions do a great job at exemplifying the strengths and possibilities offered by the approach. This dissertation builds upon these individual examples and provides a first conceptual and technical framework for augmenting social interactions.

4.5 Summary

This chapter introduced the concept of social augmentation. The overall goals of social augmentation are to increase the user's awareness of their own behaviour and improve the position of the user in the social interaction. To achieve this, six distinct requirements have been postulated, covering both the effectiveness of the augmentation and its impact on the user, the interaction and the interlocutors. Throughout the thesis, these requirements will be used to measure the degree to which theories, concepts or even systems are in line with the social augmentation concept.

At the heart of the concept lies the behavioural feedback loop which enables the analysis of user behaviour and delivery of appropriate feedback in realtime. For this, it relies on mobile social signal processing techniques and multimodal feedback generation. To increase the effectiveness of the augmentation, the chapter also introduced three mechanisms for automatically adapting the feedback loop to the user and the scenario.

Lastly, the chapter placed social augmentation in the current social skills training ecosystem. In this context, a classification scheme has been introduced which splits contemporary training approaches according to where the training happens and how the performance of the user is evaluated.

The following pages will build upon the ideas and concepts presented in this chapter, and take a close look at both behaviour analysis (see Chapter 5) and feedback generation (see Chapter 6). A concrete technical framework for social augmentation will be presented in Chapter 7, enabling anyone to design and build their own social augmentation application. Two concrete social augmentation systems are presented in the third part of the dissertation (Chapters 8 and 9).

5. Mobile Social Signal Processing

The first step in a behavioural feedback loop is the perception, analysis and classification of the user's behaviour. The results of this step will not only be used to inform the generation of feedback to be sent to the user, but also enable the system to compute the impact of this feedback. Mobile social signal processing techniques lend themselves well to this task since they enable realtime "in the wild" analysis of a wide range of user behaviours.

The term social signal processing (SSP) has been first used by Alex Pentland to describe "tools that measure social signals by analysing the statistical properties of the speaker's tone of voice, facial movement, and gesture" [Pentland, 2007]. A more general definition is provided by Vinciarelli et al., according to which "Social Signal Processing (SSP) is the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals" [Vinciarelli et al., 2009]. A social signal represents any form of nonverbal communication which bears information in a social context [Pantic et al., 2011], for example gestures, facial expressions, paralinguistic features (tone of voice, speech rate, etc.), but also less obvious behaviours such as inter-personal distance and eye gaze. Verbal features have been purposely omitted as they are not considered to be part of the focus of SSP [Palaghias et al., 2016; Pentland, 2007; Vinciarelli et al., 2009].

One aim of SSP is to enable a computer program to react to a user similarly to how a human would. For example, imagine a navigation system which would mute itself when it determines you are frustrated with it, or a robot which can tell if you are sad, in which case it cheers you up with a joke. To achieve this, SSP uses a combination of hardware and software technologies. Social signals are most commonly perceived using various types of sensors ranging from common devices such as cameras and microphones, and going all the way to complex physiological measurement arrays and motion capturing systems. The signals are processed using digital filters and lastly classified into higher level constructs depicting psychological and sociological concepts (e.g. emotions or behavioural patterns) with the help of machine learning.

Recent technological advancements have enabled the development of powerful wearable

computers, which are able to run complex computations on-the-go. The most obvious example of this evolution is the smartphone. A device which started as a simple means of peer-to-peer communication evolved into a core element of our daily lives. We use smartphones to browse the web, check emails, navigate the world, take pictures and pay in stores. I even use my smartphone to read papers, write code or configure my server. Besides the incredible computational power such devices offer, smartphones also include a variety of sensors which can measure accelerations, gravity, GPS, luminosity and even heart rate. Moreover, thanks to Bluetooth, smartphones can interface with external sensors as well, making them a true sensing powerhouse. When combining these features with the small size, large battery and ubiquitousness of these devices, we end up with ideal platforms for studying and analysing human behaviour "in the wild", e.g. during a social augmentation scenario.

While the term mobile social signal processing (mobile SSP) has been used in related literature, there is no clear definition of the concept yet. Vinciarelli et al. [2010] first mentions the term in a workshop introduction paper and places it at the crossroads between the "Mobile HCI and Social Signal Processing research communities." Palaghias et al. [2016] describes mobile SSP as a method for "extracting social behaviour on mobile devices" yet does not go into detail on the quality of the extraction or what a mobile device actually is. Thus, the following definition is put forward:

> **Definition** Mobile social signal processing is a subclass of social signal processing which does not constrain the mobility of the subject, and the data acquisition, processing and classification happens online and in realtime.

Mobile SSP usually involves the use of mobile devices, i.e. standalone small-factor computing units which the user can wear or carry around. Typical examples are smartphones, smart glasses, smart watches and smart armbands. However, approaches where data collection is done on mobile devices yet the processing is performed (fully or partially) on a remote server also fall within the boundary of mobile SSP as long as the mobility of the user is not impaired (e.g. the user can leave the room or the building). While mobile SSP enforces online processing, the training of the classification models may happen both offline (batch learning) and online (online or active learning [Settles, 2012]).

To this end, mobile SSP represents a good fit for providing behaviour analysis in a social augmentation scenario. More specifically, the mobile nature of such techniques is in line with the fifth requirement of social augmentation: The mere presence of the augmentation system has a minimal impact on the interaction and its participants (see Section 4.1). Moreover, thanks to the focus on online and realtime processing, behaviour analysis performed using mobile SSP has minimal latency and is pertinent to the moment. This allows the generation of relevant feedback which increases the likelihood of triggering a change in user behaviour (R3).

The aim of this chapter is to provide the reader with an overview of mobile SSP and illustrate the possibilities which currently exist for analysing and reacting to the behaviour of the user as part of a social augmentation scenario. But first, the chapter will offer a quick discussion on the challenges mobile SSP faces, and how they can be handled. This is followed with an overview of the different types of social signals which can be analysed with mobile SSP. In each case, examples from the literature will be provided in an effort to illustrate the state-of-the-art in terms of processing and classification approaches while also discussing the feasibility for using each social signal as part of a behavioural feedback loop. During this

literature review, only systems which satisfy the conditions of mobile SSP (as defined before) will be considered. Thus, work which focuses on offline processing or is not designed to function in a mobile setting, is not included. Finally, the chapter gives an overview of existing software frameworks for mobile SSP and discuss their fit for augmenting social interactions.

5.1 Challenges

Mobile SSP gives rise to interesting and unique challenges. While most of them are caused by the small form factor of mobile devices, some challenges also arise from the in-situ setting in which mobile SSP aims at being deployed.

5.1.1 Online Processing

Even though the ultimate goal of SSP is "building socially aware systems" [Pentland, 2007], most work of the last decade still focused on offline analysis and classification of data recorded in highly controlled environments [Wagner, 2016]. Such systems usually fail to perform as expected in live environments due to user heterogeneity and noisy data. However, for mobile SSP and social augmentation, efficient online processing is paramount. Signals need to be processed and classified in realtime to allow the system to deliver relevant feedback.

In his PhD thesis, Wagner [2016] identified several distinct set of challenges online processing faces. First, when working in a realtime scenario, inspection of the data prior to processing is not possible. This means that the system in itself needs to be able to deal with artefacts such as noisy or missing data. Moreover, in an online scenario the system no longer has access to the entire data set, but only to the data it has perceived so far, which may or may not be indicative of the user's full behavioural spectrum. Secondly, user heterogeneity might mean that not all signals fit perfectly into the predefined patterns. For example, a gesture recognizer might have been trained to detect circular hand movement yet during evaluation, one user's circles might be elongated on one axis, resembling an oval. In this case, the system needs to be able to handle such variations without impacting the overall robustness of the system.

Furthermore, one common characteristic of offline classification is manual data segmentation. Here, a person manually selects which parts of the recorded data are relevant and which not. Since in most cases this is not feasible in an online setting, the system must be able to automatically segment the data stream into relevant and irrelevant chunks. Solving this often requires the use of specialized algorithms such as dynamic time warping [Alon et al., 2009; Myers and Rabiner, 1981] or Hidden Markov Models [Narasimhan et al., 2006].

Finally, online social signal processing is more demanding both in terms of computational performance and implementation effort. Data which, in an offline setting, would be semi-automatically processed over multiple days now has to be handled fully automatically and in realtime. This is particularly critical for mobile devices where performance is already stretched thin, limiting the complexity of such a system and restricting the use of intricate machine learning algorithms. A more thorough discussion on the performance limitations of mobile devices is provided in Section 5.1.3.

5.1.2 Data Annotation

In order to enable the automatic classification of sensor data streams, a classification model first needs to be trained. During training, the model is presented with data examples of each

individual class. For this, training data needs to be acquired and labelled according to which class it is indicative of. This labelling process is also called data annotation.

The typical method for collecting training data is to perform a controlled study during which users are asked to perform different tasks. Each task is designed to elicit behaviour indicative of one specific class. For example, if the goal is to train a model for classifying stress, one task would be designed to put the user in a normal (relaxed) state, and another task would attempt to stress the user. During the data collection, a researcher would write down every time a user started and finished a particular task. Alternatively, a camera would be used to record each session and the researcher would then manually annotate the data after the study.

This process works well when the training data is collected in a controlled environment, but is unsuited for the mobile SSP context where data collection is performed "in the wild." In such scenarios, it is often not feasible to have a researcher follow each user around to see when they are stressed. Similarly, recording video or audio for post-hoc annotation is also not feasible due to memory restrictions (a data collection study can last days or weeks). Moreover, such a method would also violate the privacy of the users.

To this end, for mobile SSP, a different approach for data labelling is required. One possibility is to task the users themselves with the annotation. For this, simple annotation requests could be delivered periodically to the user's smartphone, smart watch or smart glasses. The users could then simply tap a button if they are stressed, and another one if they are not. This information could then be either stored on the device for future offline model training, or, in the case of online or active learning [Settles, 2012], it could be used to update the model in realtime.

5.1.3 Performance and Energy

Currently, in terms of computational power, the main dividing line between mobile devices and stationary computers is the processor's architecture. Stationary computers use an x86 Complex Instruction Set Computing (CISC) architecture whereas mobile devices use an ARM Reduced Instruction Set Computing (RISC) architecture. In layman terms, CISC is better suited for executing complex operations whereas RISC operations are more atomic, and thus can be better optimized for energy efficiency. From the point of view of SSP, RISC-based computers have more difficulty[1] in executing the complex operations required for some feature extraction algorithms and machine learning techniques. Besides architecture, the computational power is also constrained by the limited heat dissipation properties of modern smartphones. Whereas stationary computers are able to use active cooling solutions (air or water-based), mobile devices are restricted to less-efficient passive solutions. Thus, in order to avoid overheating, mobile processors are artificially limited to run at lower frequencies. Another drawback of mobile devices is the reduced memory size which can hinder the execution of memory intensive processing pipelines, such as those used in video analysis.

Nevertheless, thanks to the rapid improvement of computing performance over the past decade, with the current generation of smartphones offering powerful quad-core 64 bit processors and up to 6 GB of RAM, the computational power has reached a level where complex mobile SSP is feasible. Whereas in the past, mobile applications would rely on outsourcing classification tasks to a back-end server connected over network [Miluzzo et al., 2008; Muaremi et al., 2013; Rachuri et al., 2010; Tanveer et al., 2015] or resort to simple

[1] when compared with an x86 processor running at the same frequency

decision trees (DT) for classification [Wang et al., 2009; Lu et al., 2010], newer systems are able to run complex feature extractors and classification algorithms natively on mobile devices [Chang et al., 2011; Ertin et al., 2011; Lu et al., 2012]. The switch to local processing also eliminated some disadvantages associated with remote server-based processing, such as data transfer costs and privacy violations. Thus, it is also the method of choice for the present social augmentation concept.

The second constricting factor of mobile devices is energy consumption. In order for smartphones to be mobile, they draw power from an integrated finite battery, rather than a continuous power line. However, current Lithium-Ion battery technologies are very limited and, unlike processing units, have seen next to no advancements in the past decade. Although there are efforts to improving the energy efficiency of mobile devices, these mostly work by further constricting the computational capabilities of the device, and thus stand in conflict with the computational demanding nature of SSP tasks. Current smartphones are able to last roughly one day of normal use and up to 10 hours of heavy use. This is also the case for signal processing tasks. Lu et al. [2012] report a theoretical runtime of 32.6 hours when idling and 9.6 hours when classifying stress. Smaller form factor devices are less generous. For example, even when running simple pipelines, the Google Glass has an uptime of less than three hours [Gaibler, 2017].

Thus, in order to not excessively diminish battery life, algorithms running on mobile devices need to be mindful of their energy consumption. Common techniques for improving battery life are reducing sample rates [Lee et al., 2013; Rachuri et al., 2011] or gating computationally demanding processes behind simple triggers [Lu et al., 2012; Rachuri et al., 2010; Wang et al., 2009].

5.1.4 Privacy

Generally speaking, recording and analysing the behavioural data from users can be, under certain circumstances, a violation of the privacy and intimacy rights of the user. The social augmentation concept explicitly addresses this with its sixth requirement (see Section 4.1), which makes protecting the privacy of the user and the bystanders a priority for augmentation systems. Yet, in classical SSP scenarios this was less of an issue since they were usually conducted in a lab with the explicit knowledge and consent of the user. However, the switch to mobile SSP means that behavioural data can now be recorded and analysed in the wild, where not all subjects might be aware of it. Moreover, due to the small size and inconspicuousness of the sensors and processing hardware, the execution of a mobile SSP task is also far less obvious. Thus, when designing SSP pipelines for mobile applications such as social augmentation systems, it is important to be mindful of these facts and avoid recording and processing data of users which have not explicitly consented to it.

In the context of this thesis, mobile SSP is used as a component of a behavioural feedback loop where the user is both the target of the behaviour analysis and the recipient of the feedback. This means that the feedback is only derived from the user's behaviour. Thus, the privacy issue in this standard scenario is less of an issue, since the user is the one who has initiated the augmentation process, and thus should have given his consent. This ensures that only persons which receive direct benefits from the system are potentially exposed to penalties, such as privacy invasion (see Section 10.2.1 for a discussion on what happens when this restriction is lifted).

Another aspect related to privacy is data storage. For social augmentation, two types

of data storage are plausible. First, low persistency short duration storing in the device's active memory allows the system to analyse small windows of behaviour and generate feedback accordingly. Secondly, high persistency local storing can be used to allow the user to manually inspect augmentation sessions during post-hoc review phases. In this case however, to conserve memory space and protect privacy, only anonymous data should be stored. Anonymous data, although describing the behaviour of a specific user, is not sufficiently explicit to allow the identification of the user. For example, the raw audio signal is considered non-anonymous (one can identify the source person from it) whereas the acceleration data from a user's smart watch is anonymous (one cannot directly identify the person using the acceleration data). Non-anonymous data can be anonymised through processing, e.g. the energy of an audio signal is considered anonymous.

5.2 Sensors

From a conceptual point of view, a sensor is a device able to convert real world analogue signals into computer processable digital signals. Thus, for our purposes, a sensor represents an *analogue-to-digital-converter*. Through the air, information is transmitted as analogue signals, e.g. audio waves or light. A sensor converts the analogue signals into digital information through a process called sampling, during which a discrete value is assigned to the analogue signal multiple times a second. For example, a common microphone samples an audio signal by assigning a single two byte value (a short) to the signal 16000 times per second. Similarly, a common digital video camera assigns one million values (also called pixels) to the signal, 30 times per second.

Various types of sensors can be used to record behaviours. These range from common devices, such as microphones and cameras, to more specialized sensing instruments, such as physiological sensors and eye trackers. The sensors can be classified according to their physical relationship to the sensing subject into contact and remote.

Contact sensors are devices which are attached to the user's body and stand in direct contact with the user's skin or clothing. Such sensors have the benefit of allowing accurate measurement of body signals close to their point of origin. The close proximity of head-worn microphones to the user's mouth means that the recording of the user's voice is less influenced by environmental noise. The electrodes of physiological sensors allow the measurement of electrical waves along nerve and muscle fibres at skin-level to compute heart rate, muscle activity or skin conductance. Accelerometers placed on the user's body are able to track the movement of the limbs without worrying about field of view and occlusion. Despite the numerous advantages, body-worn sensors are not always the best option. The reason for this is that physically attaching objects to the user's body often times encumbers the user's movement and intrudes on the user's personal space. This can influence the user's behaviour and, in turn, the data being measured [Ouwerkerk et al., 2008]. For example, electrodes attached to a user's skin will most likely result in an increase in stress which might falsify an experiment which aims to measure user stress in various circumstances. These negative effects can be diminished by miniaturizing the sensors and embedding them in clothing (shirts, jackets), clothing accessories (glasses, armbands) or furniture [Ouwerkerk et al., 2008]. For instance, one particularly successful approach is smart armbands (also called fitness trackers) which include a wide range of sensors in a wireless armband-like device users can wear as a clothing accessory.

If intrusiveness is a large concern, remote sensors represent a viable alternative. Such sensors are characterized by their ability to perceive behaviours from a distance and by the fact that they are not attached to the user's body. One typical example of remote sensors is room microphones which can seamlessly record the voice of a user from a distant point. Depth sensing cameras are able to perform full body tracking from a distance without requiring the user to wear any markers or sensors. More advanced technologies even allow remote analysis of physiological signals [Adib et al., 2015; Wu et al., 2012]. However, the cost of this flexibility is usually a reduced signal accuracy and a higher susceptibility to environmental factors. Depth sensing cameras are plagued by noisy tracking and occlusion problems, room microphones pick up audio signals from all directions including unwanted ones, and remote physiological signal perception does not play well with moving subjects.

In a mobile SSP setting, researchers also use virtual sensors to track user activity on smartphones. This allows the collection of data such as call and messaging information, schedule information, phone usage and social media activity. However, virtual sensors suffer from poor accuracy since the behaviour of the user is only indirectly measured. For instance, the absence of appointments in the calendar does not necessarily mean the user does not have any appointments, the user might just have forgotten to add them to the calendar. Similarly, an empty call list does not mean the user does not communicate with other persons, it just means they do not do it over the phone.

Ultimately the choice of sensors depends on the scenario, the context and the users. In the case of augmenting social interactions, mobility, accuracy, robustness towards environmental factors and inconspicuousness play a front role. Therefore, mobile contact sensors are preferred. The drawbacks of contact sensors can be countered by using miniaturized and wireless devices embedded in everyday items such as clothing, glasses or watches. At the time of this writing, there is a large selection of smartphone-compatible smart armbands, smart watches and smart glasses available on the market. Remote sensors are generally less suited as they are usually static and limited to operation within specially prepared environments. Furthermore, unlike contact sensors, there are currently very few off-the-shelf solutions for smartphone-based remote sensing. Still, if the augmentation happens in fixed locations (e.g. classroom, consultation room, office), remote sensors may be used to replace body worn solutions to reduce the intrusion factor and allow more ad-hoc augmentation. Finally, virtual sensors are also a poor fit as primary drivers for social augmentation due to their slow data rate and inferior accuracy. However, they can be used as secondary data sources to provide contextual information to the current activity the user is performing.

5.3 Social Signals

Following the retrieval of raw data streams from various sensing devices, signal processing techniques are employed to analyse the behaviour of the user. The first step of this process is the extraction of behavioural cues. Vinciarelli defined the term *behavioural cue* as a "set of temporal changes in neuromuscular and physiological activity that last for short intervals of time" [Vinciarelli et al., 2008]. The duration and size of a behavioural cue can vary from a short localized event (a wink) to a longer body-wide action (a posture).

The next step is the inference of social signals from the extracted behavioural cues. Whereas behavioural cues represent just the physiological actions and are not infused with any context-specific meaning, *social signals* "concern social facts" such as "social interactions,

social emotions, social attitudes, or social relations" [Pantic et al., 2011], and thus have a specific meaning in a social context. For instance, the gesture "showing the middle finger" is first and foremost a *behavioural cue* characterized by extending the middle digit of one's hand, yet when put into the context of a social interaction, it becomes a *social signal* of contempt or dismissal.

This section will give an overview over the most common social signals analysed in mobile SSP. For each one, processing examples from related literature will be provided and their feasibility for use in social augmentation scenarios discussed.

5.3.1 Paralinguistic Signals

A review of the related literature shows that one of the most often used input signal for social signal processing is the human voice. More specifically, this involves the perception, analysis and classification of paralinguistic features.

Paralinguistic characteristics have drawn much attention from the mobile SSP community over the past decade. The most common used characteristics are energy-related features such as intensity [Boyd et al., 2016; Lee et al., 2013] and speech rate [Damian et al., 2015b; Tanveer et al., 2015]. More advanced features such as those based on frequency analysis (e.g. pitch) have also been used [Chang et al., 2011; Lu et al., 2012], however the complexity of these features represents a "computational burden" [Palaghias et al., 2016] for mobile devices. Thus, computing such features in realtime on a mobile device often involves the implementation of special mobile-friendly algorithms [Chang et al., 2011], adding dynamic processing phases which are only active when certain preconditions (e.g. voice has been detected) occur [Lu et al., 2012; Rachuri et al., 2010; Wang et al., 2009] or reducing sample rates [Lee et al., 2013; Rachuri et al., 2011].

In a mobile SSP context, researchers often use paralinguistic features to infer various higher level social signals such as emotions or stress. Chang et al. [2011] use prosodic features together with a Support Vector Machine (SVM) to achieve 75% accuracy for recognizing positive versus negative emotions, and 84% for recognizing stress from non-stress. However, while the processing pipeline was able to run in realtime on a Google Nexus One (1 GHz single-core processor) smartphone, the reported accuracy has been achieved during an offline evaluation using prerecorded semi-naturalistic corpora. Similarly, Rachuri et al. [2010] present a system for recognizing emotions using Gaussian Mixture Models (GMM) and running on a Nokia 6210 mobile phone (369 MHz single-core processor). Their approach was at that time not realtime capable, yielding recognition latencies between 50 and 140 seconds. An offline evaluation using an acted corpora yielded 71% accuracy for 5 classes of emotions. Lu et al. [2012] was among the first to attempt to take emotions classification in a more naturalistic out-of-lab environment. Their system also uses a GMM-based classifier and runs on a Samsung Galaxy Nexus smartphone (1.2 GHz dual-core processor). Using data collected in an out-of-lab environment with the help of a smartphone, they are able to achieve an offline stress recognition accuracy of 71.3%.

Paralinguistic features have also been used to detect the presence or absence of voice at a specific moment in time. This is a particularly useful classification since it can be used to gate future processing steps. That is, if there is no voice activity, there is no need to compute more complex features. Furthermore, it can also be used to compute total speaking time [Damian et al., 2016a], frequency of voiced segments [Madan and Pentland, 2006] and turn taking-related features [Lee et al., 2013]. Most commonly, the presence or absence of

speech has been computed using both energy and frequency-related features [Lu et al., 2012; Madan and Pentland, 2006]. A less computationally expensive technique is to use the zero crossing rate (i.e. counting how often the signal drops below zero) of the waveform [Lu et al., 2011; Nirjon et al., 2013].

For social augmentation, paralinguistic features present themselves as a particularly interesting option since they are good indicators for behaviour quality and, in most cases, are also fast to compute. Thus, augmentation based on paralinguistic features should be able to improve the user's position in the social interaction (R4 – see Section 4.1). However, unlike body or facial features, feedback in response to raw paralinguistic cues can be difficult to interpret. Boyd et al. [2016] noticed that upon receiving visual feedback on loudness and pitch modulation, the users experienced "difficulty in taking actions", requesting explicit "directions" in how to alter their behaviour. Thus, instructional feedback might be needed to make sure the augmentation fulfils the third requirement of social augmentation, i.e. the feedback is appropriate for facilitating a behavioural change.

5.3.2 Body Signals

Body movement has also drawn a large amount of interest over the last decade, with a great deal of it owed to the introduction of small and low-cost inertial measurement units (IMU) which can be built directly into wearable devices such as mobile phones or armbands. These sensors are able to provide accurate readings on nine degrees of freedom (three-axis accelerometer, gyroscope and magnetometer). This allows researchers to study a variety of behavioural cues by using nothing but off-the-shelf hardware.

One use case for body signal analysis is activity recognition, i.e. classifying the activity the user is engaged in at a specific moment. Simple activity recognition solutions are able to determine whether the user is moving or not [Aharony et al., 2011; Feese et al., 2013], or differentiate standing from sitting [Jovanov et al., 2013; Liu et al., 2012]. For this, acceleration or orientation data (extracted either from smartphone-internal or external sensors) is classified with the help of computationally non-intensive threshold-based solutions. Advanced approaches are able to more accurately classify the activity. Lu et al. [2010] use a decision tree on time and frequency domain features to classify standing, walking, running, driving and bicycling. Their system is able to run in realtime on an Apple IPhone (412 MHz single-core processor) and a Nokia N95 (332 MHz dual-core processor), and an offline evaluation reported an accuracy of 91.64%. A more lightweight solution is proposed by Wang et al. [2009] using only the standard deviation of the magnitudes of the acceleration. This way, the feature extraction and the classification are able to run in realtime on a Nokia N95 and using an offline evaluation they report 70% accuracy. In the social augmentation context, activity recognition can be used as a gating technique in order to determine whether the augmentation is needed or not.

Another important type of body behavioural cues is gestures. Regardless if we use gestures to convey information, to complement, disambiguate and regulate our verbal channel, or simply as a reaction to internal happenings, they play an important role during social interaction and are thus critical sources of information for social signal inference. While there is much research bearing the label mobile gesture recognition [Liu et al., 2009; Hua et al., 2007; Park et al., 2011], an extensive literature survey did not manage to find any system specifically targeting social gestures. One possible reason for this is that gestures often show "great variations among participants even for the same predefined gesture" [Liu

et al., 2009]. I argue that this effect is amplified when moving away from symbolic gestures and into the domain of conversational gestures. One solution to this is to focus on qualitative behavioural cues. Qualitative cues measure the human behaviour on a continuous scale instead of attempting to categorize it in a predefined fixed number of behavioural classes. Thus, they are akin to the qualitative features we see in other areas, e.g. audio pitch or loudness. One well defined set of qualitative body cues are the expressivity features: energy, fluidity, spatial extent, overall activation and temporal extent. They were first defined by Hartmann et al. [2005] for use in virtual agent animations based on popular psychology studies [Wallbott, 1998; Gallaher, 1992]. Baur et al. [2015] were the first to use expressivity features in a social signal processing context for their work on automatically annotating human-agent interactions. For this, they relied on a Microsoft Kinect depth camera connected to a Windows PC. On mobile phones, expressivity features can be computed using external IMUs attached to the user's arms [Damian et al., 2015a; Damian and André, 2016].

Generally, body signals are a good fit for our social augmentation scenario since feedback based on both qualitative (e.g. "your movements are too energetic", "perform larger gestures") and quantitative cues (e.g. "do not cross your arms") is easy to understand and respond to. Moreover, body signals are known to play a large role in shaping the outcome of social interactions, e.g. getting hired [Cohen and Etheredge, 1975; Drake et al., 1972; Hollandsworth et al., 1979]. Thus, they score well relative to the third and fourth requirements of social augmentation presented in Section 4.1: Feedback is appropriate for facilitating a behavioural change (R3) which is advantageous for the position of the user in the social interaction (R4).

5.3.3 Facial Signals

Signals pertaining to the facial region are also critical for social interactions. In particular facial expressions are powerful instruments in externalizing one's emotions, mood and attitude.

The most common sensor used for recognizing facial expressions on mobile devices is the video camera. One such system is Visage [Yang et al., 2012]. It uses an SVM classifier to achieve an offline recognition rate of 83.78% for seven expression classes. Whereas the approach is able to function in realtime on an Apple iPhone 4 (1.4 GHz dual-core processor) when the image resolution is very low (80 x 60 pixels), this is no longer the case for higher resolutions (640 x 480) where the face detection step for a single frame can take upwards of 4 seconds. Google's recently released Mobile Vision API[2] also offers face tracking and smile classification. While no official information on performance is available, personal small scale testing on various Android devices suggests that the face detection and smile classification is able to work in realtime with no noticeable delay. A mobile version of the popular SHORE [Ruf et al., 2011] engine has also been announced[3], however no actual implementation details or performance measurements have yet been published, suggesting limited functionality in an out-of-lab environment.

Besides video cameras, other sensors have also been used to detect facial expressions. One particularly interesting approach has been proposed by Fukumoto et al. [2013]. They detect the movement of the user's cheek using a glasses-mounted infrared sensor and, with the help of a simple threshold-based approach, are able to detect smiles and laughter with an 89.2 % accuracy. This solution is particularly interesting for our social augmentation scenario because, unlike video-based approaches, it uses a wearable contact sensor which is

[2] https://developers.google.com/vision/face-detection-concepts

[3] http://www.iis.fraunhofer.de/en/pr/2014/20140827_BS_Shore_Google_Glas.html

integrated in a normal pair of glasses. Despite the original system requiring a Windows PC for the classification, the simplicity of the algorithm should allow it to work on a smartphone. Unfortunately, the hardware part of the system is less simple, requiring a special optical sensor able to make millimetre accurate measurements.

Overall, facial expressions lend themselves well to the context of social augmentation since they are a crucial element of social interactions, and related feedback (e.g. "smile more often") can be easily interpreted and responded to by the user. Thus, facial signals score well relative to R3 (feedback triggers behavioural change) and R4 (change is advantageous for the user). The only drawback is that most approaches require a remote camera facing the user. This is problematic from the point of view of R5 since such a setup is likely to disrupt the social interaction.

5.3.4 Gaze Signals

Gaze-related cues such as eye contact or eye gaze direction offer a rich source of information for inferring higher-level social signals such as interest [Strandvall, 2009], dialogue characteristics [Mehlmann et al., 2014; Mutlu et al., 2006] or relationship [Baur et al., 2015]. Gaze cues such as saccades, blinks and fixations can be computed using a process called eye tracking. Eye tracking is usually performed on a video data stream, however alternative solutions (e.g. electrooculography - EOG) have also been proposed.

While for desktop computers there are already various commercially available solutions for eye tracking, the mobile landscape is less populated. Pino and Kavasidis [2012] propose an approach for video-based gaze detection using an optimized implementation of the Haar classifier. This enables them to compute the gaze of the user relative to the mobile phone's screen in realtime on an Apple IPhone (412 MHz single-core processor). An alternative to video-based eye tracking is offered by Bulling et al. [2009]. They use a wearable EOG system which features six dry electrodes integrated into a pair of safety goggles. Using this setup, they are able to detect various eye-movements features such as saccades, blinks, fixations and eye-gestures in an online setting with the help of a custom integrated processing unit. The commercially available Jins Meme[4] also integrates EOG electrodes, a processing unit and Bluetooth communication is an unobtrusive glass frame. However, since the device only features three electrodes and "a minimum of five EOG electrodes is required for recording eye motion" [Bulling et al., 2009], the Jins Meme is only able to detect the rough direction (up-down or left-right) of an eye movement.

Generally speaking, gaze-related cues are a good fit for the social augmentation scenarios. Providing feedback based on the presence or absence of eye contact in a social interaction could help users improve the outcome of various critical interactions such as job interview or speaking in public, fulfilling the fourth requirement of social interaction (augmentation improves interaction quality – see Section 4.1). In particular the mobility and flexibility of EOG approaches are perfectly suitable for augmenting interactions since they are less likely of disrupting the social interaction (R5). In the absence of an EOG system, motion sensors integrated in an HMD could also be used to get a rough estimate of where the user is looking. However, in order to infer eye contact from the eye (or head) orientation, a front facing camera is required. This places the approach at odds with the privacy requirement of social augmentation (R6).

[4] https://jins-meme.com/en

5.3.5 Interpersonal Distance

Measuring the interpersonal distance is an important area of research as it can be used to infer the existence of a social interaction. The most common approach to measuring the distance between two persons (each carrying a mobile device) is by using some form of radio signals such as Bluetooth or WiFi. The simplest method is to scan the environment for nearby Bluetooth or WiFi receivers [Antoniou et al., 2011; Eagle and Pentland, 2005; Miluzzo et al., 2008]. This gives a rough estimate of all devices located in a radius of 10 m for Bluetooth, or 35 m for WiFi. The resulting list can then be filtered using the Bluetooth identifier (BTID) or WiFi Service Set Identifier (SSID) in an effort to exclude devices which do not represent other persons. More advanced methods involve using the Bluetooth or WiFi Received Signal Strength Indicator (RSSI). The RSSI measures the signal strength for a connection between two devices and can be used to estimate the distance using a path loss model (PLM). PLMs are used to model the signal propagation through a specific (indoor or outdoor) environment. The simplest PLM is a free-space-path-loss (FSPL) model which assumes the free and unhindered propagation of signal through an outdoor environment. Indoor environments are more difficult, yet general PLMs for typical office environments have been proposed [Ghose et al., 2013]. To further improve accuracy, researchers have also employed machine learning techniques to classify distance from RSSI data [Carreras et al., 2012]. Besides Bluetooth or WiFi, audio signals can also be used to determine the distance between two devices (one emitter and one receiver) by measuring the time it takes the audio signal to travel between the devices [Peng et al., 2007; Xu et al., 2011].

Proxemics can play an important role during social interactions, especially in multi-cultural encounters where researchers found large variances in interpersonal distances [Barnlund, 1975; Ferraro, 1990]. In such scenarios, social augmentation techniques could be used to minimize cultural misunderstandings and generally improve the quality of the social interaction (R4 – see Section 4.1). Proxemics analysis can also help the system automatically adapt the behavioural feedback loops to the current interaction context (see Section 4.3.1).

However, looking at the work done so far in mobile SSP related to interpersonal distance classification, the available techniques lack the necessary accuracy for detecting the fine variations in interpersonal distance which often mean the difference between appropriate and inappropriate behaviour. Moreover, the fact that for RSSI based proximity classification all potential interlocutors are required to carry specialized hardware or smartphones with pre-installed software makes it a less than ideal solution for “in the wild” augmentation.

5.3.6 Physiological Signals

Physiological signals have also gathered the attention of researchers and engineers. These cues allow the measurement of normally unperceivable happenings within the user and can be used to infer various high-level social signals such as emotions or stress. The most common physiological cues used in mobile SSP are heart rate, electrodermal activity, skin temperature, breathing rate and brain activity. Heart rate is usually measured with the help of electrocardiography (ECG) [Ertin et al., 2011; Gaggioli et al., 2012; Gluhak et al., 2007], which involves the use of skin-level electrodes to measure electric changes in the user’s body. Alternatively, heart rate can also be measured using photoplethysmography (PPG) [Poh et al., 2010] — i.e. optical sensors which measure changes in light absorption of blood vessels. PPG

is commonly found on smart armbands such as the Empatica E3[5] or the Microsoft Band 2.[6] Electrodermal activity (EDA) [Ertin et al., 2011; Gluhak et al., 2007], also known as galvanic skin response (GSR) or skin conductance (SC), measures the resistance or conductance of electricity at skin level. The electrical properties of the skin varies with the activity of the sweat glands, making EDA a good predictor for physiological arousal. Breathing rate is usually measured using a respiratory inductive plethysmograph (RIP) [Ertin et al., 2011; Gluhak et al., 2007], which requires the user to wear a stretchable band on the chest. Skin temperature is measured using electrical temperature sensors placed in the vicinity of the skin [Gluhak et al., 2007; Hu et al., 2012]. Finally, brain activity measurement relies on electroencephalography (EEG) [Campbell et al., 2010] and involves the placement of wet or dry electrodes on the head of the user.

One use case for physiological signals is the recognition of stress. Ertin et al. [2011] proposed a system able to extract and analyse the user's EDA, breathing rate and heart rate using a body worn sensor suit connected to an unspecified android smartphone over Bluetooth. Their system uses SVM or decision tree-based classification of ECG and RIP features to achieve offline stress recognition rates between 80% and 90% [Plarre et al., 2011] for two classes. However, when the data is analysed online on the smartphone, the authors report a processing delay of 1-2 minutes. Carbonaro et al. [2011] and Gaggioli et al. [2012] proposed a similar setup, but combine physiological measurements with activity-related features and use Artificial Neural Networks (ANN) for classification. However, no accuracy or performance measurements have been reported. Moreover, unlike Ertin's system, the processing is done remotely on a server and only the data collection has been performed using mobile phones.

Physiological signals have also been used to analyse the emotional state of the user. Gluhak et al. [2007] make use of ECG, EDA, skin temperature and breathing rate features to classify valance and arousal. In an offline scenario, they achieve recognition rates between 70% and 80% using an SVM classifier. For online analysis on mobile phones they replace the SVM with a less computationally intensive threshold-based approach which is able to detect two classes of arousal (relaxed and activated), yet no accuracy measures have been reported.

Despite their accuracy, physiological signals are more difficult to implement in a social augmentation scenario. Designing feedback in response to them is challenging since users lack direct control. For example, in a stressful situation, informing the user that their heart rate is too high is not that effective as they might not be able to do anything about it. Thus, feedback based on physiological signals might not be able to facilitate a behavioural change from the user (R3). However, it is possible to envision more advanced feedback routines (e.g. meditation techniques) which do not directly attempt to address the physiological signals but indirectly try to guide the user to a different (e.g. relaxed) mental state.

5.3.7 Virtual Signals

We use smartphones for a variety of activities including various forms of communication, work and entertainment. This usage generates a large amount of data which can also act as input for signal processing. To collect this data, researchers use "virtual sensors" which track and log a user's activity on a smartphone.

Phone activity data can be used to infer higher level behavioural states such as emotions or stress. For instance, LiKamWa et al. [2013] use phone activity features in an effort to infer the

[5] https://www.empatica.com

[6] http://www.microsoft.com/microsoft-band

valance and arousal of the user. Using long-term data histograms for email, SMS, phone call, internet and application activity as well as location data, they are able to classify daily valance and arousal with 93% accuracy in an offline user-dependent evaluation. Muaremi et al. [2013] combine phone usage features with audio, breathing and activity features. They achieve a user-dependent offline recognition rate of 61% for three classes of stress (low, moderate and high). For both systems, the processing is distributed between a smartphone for collecting data and a webserver for processing and classification.

In the context of social augmentation, virtual signals are a poor fit since they are seldom pertinent to ongoing social interactions. Thus, feedback based solely on the user's virtual signals is unlikely to inspire a behavioural change which would impact the position of the user during a social interaction (R4 - see Section 4.1). However, virtual signals could be used to infer the characteristics of the social interaction, for example the relation between the user and the interlocutors. Such information can be used to adapt the augmentation to better suit the interaction the user is engaged in (see Section 4.3.1).

5.4 Existing Frameworks

This section will explore existing mobile SSP solutions and discuss their fit for the social augmentation scenario. For the reader's convenience, all systems are listed in Table 5.1. As a pre-selection criteria, the section will only list and discuss systems which fit the mobile SSP definition, i.e. systems which do not degrade the mobility of the user and which are capable of online processing. Furthermore, the overview only includes frameworks which offer some form of signal processing functionality. Pure data collection systems such as MyExperience [Froehlich et al., 2007], Lathia's "Libraries for Computational Social Science" [Lathia et al., 2013], Funf [Aharony et al., 2011] or the IoTool[7] have been excluded from the overview.

One of the first social signal processing framework to support mobile devices is the Context Recognition Network (CRN) Toolbox [Bannach et al., 2008]. CRN can be deployed on both mobile (IPhone, Nokia) and desktop (Windows, Linux) computers. It offers a wide range of sensing, processing and classification components which can be combined to achieve multimodal activity and context recognition systems. Similarly, the BeTelGeuse [Kukkonen et al., 2009] is also an open source social signal processing framework specifically targeting mobile devices. The framework is able to interface with both device internal and bluetooth-connected sensors and offers advanced processing methods as well as support for SVM classification. Another example is the FieldStream [Ertin et al., 2011] framework. FieldStream is written in Java and has been implemented for the Android environment. It provides support for standard android sensors, and for the bluetooth-connected AutoSense [Ertin et al., 2011] and Zephyr[8] sensors. Moreover, it features a wide array of features and predefined models for stress and conversation recognition. However, active development has ceased for all three frameworks. Considering the extremely dynamic smartphone landscape, software older than two or three years is likely to have limited functionality when running on modern devices. Moreover, older frameworks most of time rely on no longer available sensing hardware and lack any support for modern peripherals (e.g. smart armbands). This makes active development a crucial characteristic for reusability. One example for an actively-developed

[7] `https://iotool.io`
[8] `https://www.zephyranywhere.com`

Framework	Sensing/ Processing	Classification	Platforms	Interface	Open Source	Latest Release
CenceMe [Miluzzo et al., 2008]	audio, activity	DT	Symbian	Java		2008[3]
BeTelGeuse [Kukkonen et al., 2009]	activity, physio., phone usage	SVM	Windows, Linux, MIDP[1]	Java	✓	2009
EEMS [Wang et al., 2009]	audio, activity	DT	Java	Symbian		2009[3]
FieldStream [Ertin et al., 2011]	activity, physio.	SVM	Android	Java	✓	2010
Jigsaw [Lu et al., 2010]	audio, activity	DT, GMM	Symbian, iOS	Objective-C, C++		2010[3]
EmotionSense [Rachuri et al., 2010]	audio, activity	GMM	Symbian	Python		2010[3]
SociableSense [Rachuri et al., 2011]	audio, activity	GMM	Symbian	Python		2011[3]
Auditeur [Nirjon et al., 2013]	audio	DT, NB, GMM, SVM, HMM, kNN	Android	Java		2013[3]
CRN [Bannach et al., 2008]	audio, activity, physio.	HMM, kNN	iOS, Symbian	C++, GUI	✓	2015
Dynamix [Carlson and Schrader, 2012]	audio, activity, physio.	-	Android	Java	✓	2016
SSJ (Section 7)	audio, video, activity, physio.	NB, SVM, ANN, other[2]	Android	Java, GUI	✓	2017

[1] Mobile Information Device Profile, a discontinued API supported by a variety of Nokia phones

[2] can interface with OpenSSI [Wagner et al., 2013] for classification

[3] not publicly available

Abbreviations: Artificial Neural Network (ANN), Decision Tree (DT), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), k-Nearest Neighbours (kNN), Naïve Bayes (NB), Support Vector Machine (SVM)

Table 5.1: Overview of existing frameworks for mobile social signal processing.

SSP system is Dynamix [Carlson and Schrader, 2012]. The Android-based open-source framework is able to infuse third-party applications with signal processing functionality using a dynamic plugin-based architecture. Yet, Dynamix only offers basic sensing and processing functionality (e.g. computing heart rate or sound pressure level) and lacks support for signal classification.

There are also some closed source solutions. One such example is the CenceMe system by Miluzzo et al. [2008]. It offers audio, accelerometer, GPS and Bluetooth data processing as well as discriminant analysis and decision tree-based classification. The entire system runs in Nokia's Symbian environment, more specifically on the Nokia N95. The concept has been picked up by other researchers which extended it with dynamic processing pipelines to improve energy efficiency [Lu et al., 2010; Wang et al., 2009] or allow the system to automatically adapt to context [Rachuri et al., 2011, 2010]. In contrast, the Auditeur [Nirjon et al., 2013] platform specializes on sound recognition and provides a more comprehensive range of audio features as well as various classifiers including GMM, HMM and SVM. Similarly to the work done by Rachuri [Rachuri et al., 2011, 2010], Auditeur also offers an intelligent management of processing pipelines. What sets Auditeur apart is that the system automatically downloads custom processing pipelines matching the current context and applies them in realtime. Unfortunately, all these systems are not only closed source but also publicly unavailable, making it impossible to use them in our scenario.

In summary, while there are several mobile SSP solution available, they lack the multimodal signal processing functionality required for our social augmentation endeavour, are incompatible with modern mobile devices and sensors due to a lack of active development, or are simply unavailable. In particular, none of the examples above are able to work with highly wearable devices such as the Google Glass or smart watches. Moreover, most of the above mentioned solutions lack a flexible and accessible way for creating mobile social signal processing pipelines, demanding either in-depth programming expertise or an elevated understanding of the underlying system architecture. This restricts the user base of such systems to computer scientist or technically-aligned social scientists. However, for the proposed social augmentation scenario, it is important that the users themselves are able to adapt the system and tailor the overall experience to their needs and desires. To address these issues, as part of this dissertation, a custom solution for online social signal processing fully compatible with the goals and requirements of social augmentation has been implemented (see Section 7).

5.5 Summary

Social signal processing has come a long way. Throughout the past decades we have seen the birth of ever more complex processing and classification algorithms. However, the greatest advancement has been the shift from artificial laboratory studies to naturalistic "in the wild" processing. At the heart of this shift lie the smartphones — small, mobile, powerful and, most of all, ubiquitous computers.

Mobile SSP describes the process starting with the extraction of behavioural cues from sensor data and ending with the inference of higher level social signals and social behaviours. Yet, it does this unobtrusively, without intruding on the user or the social interaction the user in engaged in. This is in line with the fifth requirement of augmenting social interactions (see Section 4.1), making mobile SSP a critical component of the overall concept.

Throughout the chapter, an extensive overview of the current state-of-the-art for mobile

analysis and classification of various social signals has been presented. From this overview, the ability of mobile SSP to facilitate the understanding of not only what the user is doing, but also why becomes apparent. This characteristic allows augmentation systems using mobile SSP to accurately choose feedback which is relevant to the user, the moment and the situation – fulfilling R3 and R4.

Yet, from an implementation point of view, while there are already multiple solutions available, none offer the flexibility and functionality required by the proposed social augmentation concept. To this end, a novel software framework has been implemented which is specifically designed for augmenting social interactions (see Section 7).

6. Live Feedback

The second stage of the behavioural feedback loop handles the generation and delivery of feedback events to the user. This is done in accordance to the results of the behaviour analysis as well as user and scenario-specific context information. However, what, how and when to send to the user is not as easy to answer. While preparing to write this dissertation, I firmly believed that there is a golden rule of feedback delivery, and I was set on finding it. This was supposed to be a silver bullet in live feedback strategies which would allow a system to insert information into the user's mind while barely (if at all) disturbing the primary task in any scenario and for every user. However, the more I read about cognitive psychology and the more I studied how humans react to feedback during real social interactions, I came to the conclusion that there is no golden strategy for feedback delivery just the same as "there is no best modality [for notifications]" [Warnock et al., 2011]. Some feedback strategies will work better in some situations, other strategies in other situations. For example, visual feedback might be more suited for public speaking scenarios as the larger distance between the user and the interlocutors is more forgiving with short eye contact breaks, however it might break social conventions during a face-to-face interview. Yet, even within a single scenario, different users can react differently to feedback. For example, for some persons, breaking eye contact with the audience during a presentation might be unimaginable whereas others have no problem doing it even during face-to-face conversations.

However, this flexibility in live feedback design has not yet been explicitly studied. This becomes apparent when looking at the current state-of-the-art. Most systems which make use of live feedback use different feedback strategies without a compelling reason to do so, suggesting a certain degree of confusion when it comes to feedback design. Boyd et al. [2016] opted to use textual feedback for one system mode and symbolic feedback for another. Moreover, both designs were visual despite arguing that the visual channel is "overcrowded" during social interactions. Schneider et al. [2015] used two different feedback strategies depending on the severity of the behavioural "mistake" the user does. The *action* feedback consists of a visual message with both textual and symbolic components as well

as a single short vibrotactile feedback. The *interruptive* feedback also contains visual and tactual components, but the visual component is more obtrusive (larger) and it also adds an auditory component (a "pause" sound). Tanveer et al. [2015] also opted for visual feedback, implementing two strategies: one symbolic and one textual. However, the textual feedback was instructional in nature (provided instructions on how to adapt the behaviour) and the symbolic one appraisive (gave information on the current state of the user's behaviour). The sheer heterogeneity of designs reported in related literature makes comparing the reported results difficult. For example, while empirical findings reported by Tanveer et al. [2015] suggest that the textual instructional feedback was better received by the users than the symbolic appraisive one, it is not clear whether the information representation or the scope of the feedback caused the difference. Furthermore, different authors use different terminology to describe feedback, further adding to the confusion.

To get a grip on this problem, this chapter will thoroughly explore the design space of feedback in social augmentation scenarios, presenting the reader with a taxonomy of live feedback. More specifically, this section will classify feedback according to their modality (visual, auditory, tactile, thermal, olfactory, gustatory or multimodal), duration, prominence, scope (appraisive or instructional) and level of detail.

This is a similar endeavour to Bernsen's work on the modality theory [Bernsen, 1994] which attempted to clearly define and classify the various input and output modalities from a designer's point of view. However, analogue to the problem encountered by Bernsen, the sheer number of combinations of feedback strategies, social behaviours and scenario parameters makes answering the question of which feedback to use under which circumstances a similarly "intractable theoretical problem." This intractability is worsened by the inherent complexity and heterogeneity of the human nature. Still, to provide the reader with practical insights for designing social augmentation systems, this chapter will use examples from related literature and psychological theories to discuss the advantages and disadvantages of every feedback dimension relative to the requirements of social augmentation introduced in Section 4.1. This should give the reader a clear picture of what options exist for delivering live feedback as well as how these options can impact the social augmentation.

6.1 Modalities

We rely on our senses to navigate through life. However, we use each sense differently, according to their capabilities and limitations. For example, our sense of vision is accurate in perceiving spatial information and extracting details from a distance, whereas our auditory sense is good for perceiving temporal information [Freides, 1974; Nesbitt, 2003; Sigrist et al., 2012]. Knowing and using these limitations is important and "designs of augmented multimodal feedback should exploit the modality-specific advantages." [Sigrist et al., 2012].

This section will provide the reader with an intuitive yet, for a social augmentation scenario, exhaustive overview of feedback modalities. For each modality, their advantages and disadvantages will be presented using examples from literature and their feasibility in our social augmentation scenario discussed. An empirical comparison between modalities in a social augmentation scenario is provided in Chapter 9.

According to Bernsen [Bernsen, 2008], a modality is defined by the physical medium over which the information is transmitted (sight, touch, hearing, smell, taste) and the way the information is represented. For the sake of readability, this section will first and foremost

Figure 6.1: The Google Glass HMD.

classify feedback methods according to the physical medium over which the information is communicated. The different types of information representation will be handled in the individual modality sections. Moreover, since the feedback is meant as a form of communication in a computer-*human*-interaction scenario, the classification scheme will closely mimic the classification of the human sensory system and ignore any modalities not perceivable by humans.

6.1.1 Visual Feedback

The most common type of feedback delivery is using the visual channel. It is being heavily used in both human-human and human-machine interactions. In technology-enhanced systems, visual feedback is commonly referred to as "secondary display." The term combines the auxiliary nature of the feedback relative to a primary task, with the word display, which symbolizes the visual nature of the feedback. The most frequent types of visual feedback systems are notification systems that alert the user to various events of background tasks (e.g. incoming email, available updates) through the use of pop-ups, i.e. small temporary windows which occlude part of the display. Visual feedback has also been used in behaviour training systems. Barmaki and Hughes [2015] provided symbolic visual feedback to help teachers improve their use of postures in the classroom. For the TARDIS system (see Section 3.1), we used icons to give the user information on which social cues have been detected during a job interview simulation. With the appearance of lightweight wireless head-mounted displays such as the Google Glass (see Figure 6.1), visual feedback can be delivered ubiquitously with only minor drawbacks. For example, both Tanveer et al. [2015] and Boyd et al. [2016] experimented with delivering multiple types of visual feedback using the Google Glass to provide information on loudness, pitch or voice modulation in out-of-lab environments. A similar system will also be introduced in Section 8.

The main advantage of visual feedback comes from the humans' ability to extract complex information from visual messages in a relatively short time. For example, we can quickly glance at an object and extract its colour, shape, pattern, position and orientation. However, if we were to describe the same object to another person using the auditory channel only, it would take much longer to communicate these properties. This is backed up by research which suggests that vision is the most accurate modality for assessing position [Pick et al., 1969; Warren and Cleaves, 1971], shape [Abravanel, 1971; Jones, 1981; Miller, 1972; Rock

and Victor, 1964] and size [Milewski and Iaccino, 1982; Seizova-Cajić, 1998], as well as allowing for faster information retrieval [Jones and O'Neil, 1985]. User interface designers have taken advantage of this by primarily using visual elements (e.g. icons, animations, labels, windows) in the user interfaces of computer applications.

One disadvantage of visual feedback is that we usually need to physically change focus in order to perceive it, resulting in a gaze interruption from the main task. The more complex the feedback, the longer the interruption. Particularly in a social setting, such gaze interruptions are known to have a negative effect on the interaction [McAtamney and Parker, 2006]. The impact of physical attention switching add to the already disrupting effects of distributed attention management when dealing with competing tasks (see Section 2.2). Kosmalla et al. [2016] found that during highly visual tasks such as climbing, visual feedback is inferior to auditory or tactile feedback, with users reporting a decrease in climbing performance.

To diminish these interferences, systems could deliver feedback to the user's peripheral field of view. This would eliminate the need to shift focus in order to perceive the feedback. Recent findings show that some stimuli which are presented peripherally are processed by different parts of the brain than stimuli presented centrally [Bayle et al., 2009]. This would suggest that peripheral vision might demand other cognitive resources than central vision. Most commonly, the peripheral field of view represents everything outside the central field of view (foveal vision). The foveal vision represents 5° of the field of vision [Millodot, 2014], stretching to either side from the centre (from $-2.5°$ to $+2.5°$). However, peripheral vision is known to be less accurate than central vision:

> "When I look at something it is as if a pointer extends from my eye to an object. The 'pointer' is my gaze, and what it touches I see most clearly. Things are less distinct as they lie farther from my gaze. It is not as if these things go out of focus – but rather it's as if somehow they lose the quality of form"
>
> (Jerome Y. Lettvin [Lettvin, 1976])

One main source of quality loss in the peripheral vision corresponds to a decrease in colour sensitivity due to a reduced density of cone receptor cells [Abramov et al., 1991; Curcio et al., 1990]. It has been reported that the vision becomes dichromatic beyond 30° and full monochromatic beyond 60° [Wooten and Wald, 1973]. Nevertheless, studies have shown that shapes and patterns can be recognized using peripheral vision [Strasburger et al., 2011]. In human-computer interaction (HCI), peripheral vision is often used in ambient information systems (see Section 6.2.1).

Another disadvantage of visual feedback is its reliance on the presence of a display in the user's field of view. Typical options are stationary monitory or body worn displays. Both options have benefits and drawbacks. Whereas stationary monitors are non-intrusive but lack mobility, body worn displays such as head mounted displays (e.g. Google Glass — see Figure 6.1) are able to deliver feedback on-the-go but their presence can interfere with the interaction. More specifically, HMDs can cause interferences from a physical point of view (e.g. size and wires restrict movement and impact vision) but also from a social point of view (HMDs can be considered "creepy and rude" and break social convention[1]).

In terms of visual representation of information, two general classes of visual feedback can be found in feedback-systems: textual and symbolic. Literature suggests that symbolic

[1] `https://sites.google.com/site/glasscomms/glass-explorers`

feedback is superior to textual feedback in scenarios such as those found in social augmentation. More specifically, Larkin and Simon H. [1987] argue that there is a difference in effort associated with making inferences between textual and pictorial information representation. According to "resource pool" types of attention models, distinct types of activities draw capacity from different "pools" [Navon, 1984]. Since a social interaction involves a large amount of processing of verbal messages, it is plausible that symbolic feedback would cause less interference with the primary task (R2 – see Section 4.1) than text-based feedback as it uses capacity from a different "resource pool" (image recognition versus language processing). However, there is no concise empirical evidence for a performance advantage of pictorial interfaces over text-based interfaces (or the other way around) [Benbasat and Todd, 1993]. Nevertheless, symbolic representations also benefit from being more recognisable than text when rendered at smaller sizes [Kline et al., 1990]. This is particularly relevant when working with small display HMDs such as the Google Glass.

Overall, visual feedback presents itself as a good fit for social augmentation scenarios. The commonness of the medium as well as the high information bandwidth means that visual feedback can easily satisfy both R1 (feedback is perceivable) and R3 (feedback is sufficient to facilitate behavioural change). Still, it is to be expected that textual feedback is more likely to satisfy R3 due to its descriptive nature. However, the fact that users need to physically change focus to perceive the feedback can impact the attention level dedicated to the primary task (R2). This problem is strengthened by the involuntary attentional capture effect which automatically directs focus towards novel stimuli (e.g. visual feedback events) in the field of view. Moreover, the reliance on displays (either head mounted or remote) can disturb social protocol and may raise privacy issues, violating R5 and R6. Peripheral visual feedback represents a special case with opposite qualities: The subtleness may result in a reduced impact on the primary task (R2) but the loss of perception accuracy might conflict with R1 and R3.

Comparing between Visual Feedback Methods

To explore what type of visual feedback shows most promise in a social augmentation scenario, we conducted a small scale user study where we compared between different visual feedback designs [Awadeen, 2015].

Study Design.

A total of 10 students, 9 male and 1 female with a mean age of 28.6 took part in the study. The participants were asked to type a text on a computer which was read to them by a different person. While typing, the users received formatting instructions via a Google Glass. More precisely, the Google Glass instructed the user to either use or not use capital letters. Each participant did this four times, once as a control and then once for each of the three visual feedback methods described below. The order of the conditions was randomized. In the control condition, the Google Glass was switched off and the capitalization instructions were provided verbally by the second person.

In this scenario, listening and typing down the text was considered the primary task whereas the formatting instructions provided via the Google Glass was the secondary (visual feedback) task. This study setup allowed us to simulate a dual task scenario while also enabling to objectively measure the impact of the various feedback methods on the user.

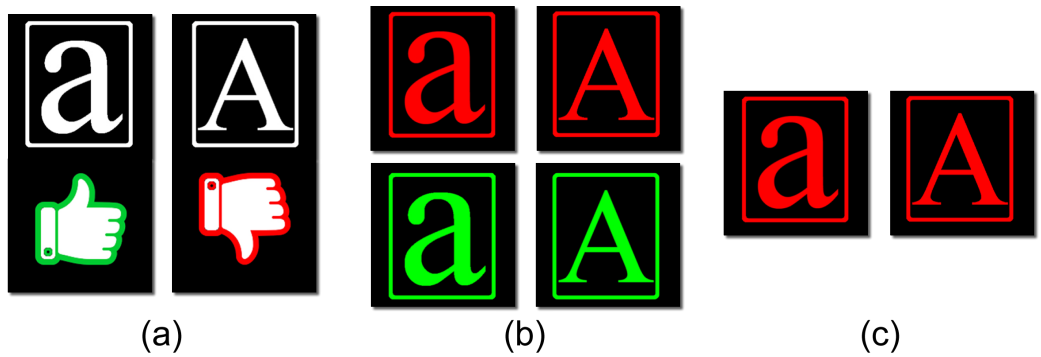

Figure 6.2: The evaluated visual feedback methods: (a) double icons, (b) coloured icons, (c) fading icons.

Feedback Methods.

Three feedback methods (double icons, coloured icons and fading icons — see Figure 6.2) have been implemented. To be in line with R3 (feedback is sufficient to facilitate behavioural change — see Section 4.1), each method conveys two distinct pieces of information: (1) the current state of the user's behaviour and (2) whether this state is appropriate or not. All methods have been built with mobile behaviour augmentation in mind. Thus, they have been optimized for use on the small screens of popular HMDs such as the Google Glass.

The double icons method (see Figure 6.2-a), consists of two icons. One shows the current state and the second icon provides information on the appropriateness of this state. The first icon was designed to resemble the caps-lock symbol with either an upper or lower case *A*, depending on the current state. The second icon consisted of a green thumbs up, displayed when the current state is in line with the system's formatting instructions, or red thumbs down sign when the user needs to change the current caps state. Both icons were persistent, i.e. always visible.

Whereas the double icon was designed to explicitly avoid colour-only information encoding in an attempt to counter HMD's poor colour contrast, the second method (see Figure 6.2-b) moved past this precaution, combining the two icons into one single colour coded icon. Similarly to the first method, the feedback was persistent, and the icon was designed to resemble the caps-lock symbol. The appearance of the symbol changed according to the keyboard's current caps state. Unlike the first method, instead of a second icon to display the appropriateness of the state, the icon itself would change colour between green (current state is good, no change required) and red (current state is bad, change is required).

Finally, the third feedback method (see Figure 6.2-c) employs the metaphor that the user only needs to be informed of a "bad" behaviour. To this end, only when the current state does not match the requested formatting does a red caps-lock symbol (either lower or upper case, depending on the current caps state) become visible. If the current state is in line with the one requested, no icon is visible. To minimize the distraction effect of the method, the icon gradually blends in and out instead of "popping" in or out.

Results and Discussion

To measure the effectiveness of the feedback we computed the time delay between the display of a visual feedback (i.e. caps on or off) and the moment when the user pressed the caps

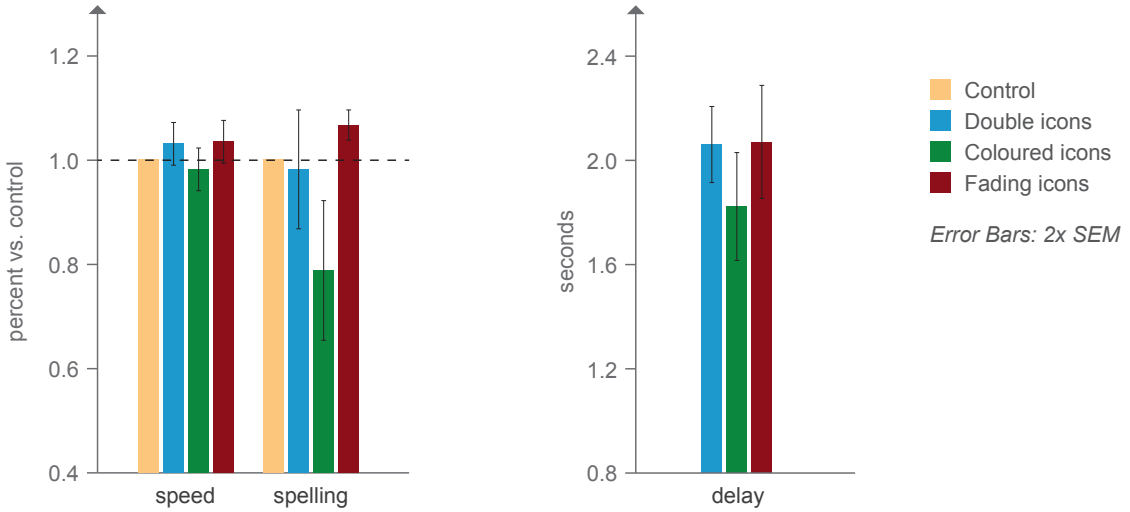

Figure 6.3: Results of the data analysis.

lock key. To measure the disruption effect, we calculated two typing quality features: typing speed and spelling (similarity with the source text). The mean values over all participants are illustrated in Figure 6.3. Due to the low number of participants, no significant differences were found. However, when looking at the spelling dimension, we can observe a trend. More specifically, the fading icons scored best (6.66% improvement over control) and the coloured icons worst (21.3% degradation over control), despite the conditions being randomized.

Due to the lack of any significant differences, the most obvious conclusion is that all three methods had a similar impact on the user and the task. Furthermore, no feedback method was significantly different from the control condition suggesting a negligible distraction effect as well as a high feedback response rate. To this end we can say that icon-based visual feedback fulfils the requirements of the behavioural feedback loop (see Section 4.1). In particular, the study shows that visual feedback is able to meet the first three requirements: feedback is able to draw some attention (R1), feedback causes a minimal disruption of the primary task (R2) and feedback facilitates behavioural change (R3).

Still, the study yielded some interesting trends. More precisely, the fading icon appear to be the least disturbing feedback method, whereas the coloured icons the most (Figure 6.3). This may be explained by the fact that only showing negative feedback reduces the overall amount of feedback messages the user has to perceive and interpret. Furthermore, the rather poor contrast of optical see-through HMDs appears to make the perception and interpretation of the coloured icons more difficult, which is in line with related literature [Gabbard et al., 2010].

Another interesting effect the study revealed is the rather long reaction time to the feedback. Regardless of the feedback method, the users needed around two seconds to press the caps lock key. While this is not necessarily a critical aspect for typical behavioural augmentation scenarios, it does suggest that the cognitive structures for perception and interpretation of visual stimuli might have been overloaded due to the visual nature of the typing task. Thus, it is plausible that another modality (e.g. tactile) might have been able to elicit quicker response times.

6.1.2 Auditory Feedback

Auditory feedback is also a heavily used mechanism for feedback delivery. Verbal feedback has been used in human-human interactions long before the appearance of computers. On a biological level, auditory feedback, in the sense of listening to our own voice, is an important aid in speech production [Fairbanks, 1955]. Toddlers also use it to control the learning of speech items as they try to imitate another person's sounds [Brainard and Doupe, 2000]. Nonverbal auditory feedback, while not as common in human-human interactions, is a very important aspect of our interactions with inanimate objects. For example, knocking on an object tells us whether it is hollow, or the sound of a car engine tells us whether it is running smoothly.

In human-machine interaction scenarios, auditory feedback cues are a very common form of abstract feedback. Everybody has at least a dozen devices which beep, buzz or humm. Devices such as mobile phones, washing machines and cars rely on auditory feedback to inform the user of an event or a change in status. One of the reasons for its ubiquitous is its simplicity: A single beep from a less-than-a-dollar speaker is enough to inform the user of an incoming call, that a washing cycle has finished or of the distance to the garage wall. The same principles could be used in behavioural feedback loops as well, for example a beep when the user's voice is too silent, double-beep when it's too loud. A concrete example for this will be presented in Section 8.

Sound has also been used to supplement visual information delivery through the use of auditory icons. According to Gaver [Gaver, 1986], "an auditory icon is a sound that provides information about an event that represents desired data." Such auditory icons alert the user to the occurrence of an event, e.g. incoming email, but can also provide additional information, e.g. a large email produces a stronger sound than a small email. A similar approach has also been proposed by Blattner et al. [1989]. However, whereas Gaver mostly focused on reusing natural occurring sounds, Blattner's "Earcons" follow the design of traditional visual icons and can be either representational (similar to a natural-occurring sounds) or abstract.

A significant amount of research has also been done in the area of sonification, i.e. the representation of information using the acoustic channel [Kramer et al., 1999]. One application for this approach is sensory substitution e.g. for visually impaired persons. Such approaches attempt to communicate information to the users on the acoustic channel for which normal sighted persons would use vision (e.g. colour [Bologna et al., 2009; Banf and Blanz, 2013], depth [Dietz et al., 2016; Twardon et al., 2013], facial expressions [Dietz et al., 2016]).

The largest advantage of auditory feedback is the fact that it can be used in parallel to the visual channel. In our day to day lives we are used to combining information from both channels into a single mental model. For example, Dodd [1977] showed that during speech perception, we rely on both hearing and vision. This makes auditory feedback especially suitable when the primary task is highly visual.

However, in case the primary task also involves acoustics, for example during a conversation, a study by Ofek et al. [2013] showed that audio feedback leads to an increased amount of distraction. More precisely, they studied the effect of visual and auditory secondary tasks during one-on-one social interactions. They found that the second user was more likely to notice when the first user received feedback, if the feedback was auditory rather than visual. This is backed up by earlier experiments where it was found that when concentrating on one audio source, we are unaware of the content (although not of the presence) of a secondary audio signal [Cherry, 1953]. These findings are in line with various attention theories which

predict that concurrent perception of similar signals can result in interferences and cognitive bottlenecks (see Section 2.2).

Similarly to visual feedback, information in auditory feedback can also be represented either verbally (using spoken words) or more abstract (using auditory icons or earcons). The largest benefit of an abstract representation is that it allows the feedback to be simple and short, making it quicker and easier to be perceived by the user. For example, an instruction to smile more might be represented as two auditory beeps, whereas smile less as one beep. The drawback of abstract feedback is that it does not provide a natural mapping between the feedback and the behaviour it relates to. This can cause the user to misinterpret the meaning of the feedback.

From the point of view of the presented social augmentation concept, auditory feedback can be a good option. Yet, especially regarding R2 (feedback does not take too much from the attention dedicated to the primary task — see Section 4.1), careful design considerations are required to limit the disruptive power of audio signals. Moreover, auditory feedback is also susceptible to environmental factors such as background noise, potentially impacting the perceptibility of the feedback (R1).

6.1.3 Tactual Feedback

The sense of touch plays a crucial role in our day-to-day lives, enabling accurate and rapid interaction with the world around us. Furthermore, it allows us to quickly gather information from physical objects, such as solidity, texture and temperature; and also plays an important role in social interactions where it can communicate emotions [Hertenstein et al., 2006].

The exact definition of tactual perception is somewhat fuzzy as the term is often misused and confused with haptic or tactile perception. According to Loomis and Lederman [1986], tactual perception, or more commonly known as the sense of touch, comprises two distinct modalities: tactile and kinesthetic perception. Tactile perception (also called cutaneous sense) provides "awareness of stimulation of the outer surface of the body by means of receptors within the skin." On the other hand, kinesthetic perception refers to the "awareness of [the] body posture" such as the relative position of the head or limbs. Haptic perception is the simultaneous perception of both tactile and kinesthetic stimuli [Loomis and Lederman, 1986].

The most common use of tactual feedback found in human-computer interaction is vibration feedback. For example, our phones vibrate when we get a call or our game controller vibrates when we perform an action inside a game environment. When looking at training and assistive technologies, vibrotactile feedback has been used as a replacement for visual and auditory feedback when these perception channels are either unavailable or otherwise engaged. Susanne Boll and her team explored the use of vibrotactile belts to help visually impaired navigate the environment [Henze et al., 2006], or as a hands-free alternative to visual directions for pedestrian [Pielot et al., 2010, 2011] and car navigation [Asif and Boll, 2010]. Vibrotactile feedback has also been proposed to help users with their presentation skills [Damian and André, 2016; Schneider et al., 2015] or improve their violin playing techniques [van der Linden et al., 2011b].

The main advantage of tactual feedback for the social augmentation scenario is that social interactions primarily involve the visual and auditory channel. This makes tactual feedback an ideal candidate to accompany such primary tasks as it is less prone to structural interferences, and thus less likely to disrupt the primary task. However, so far this effect remains theoretical as there is little practical evidence of it in related literature. Studies in

automotive and pedestrian navigation did not find any significant differences between tactile and visual displays in terms of cognitive workload and distraction [Asif and Boll, 2010; Pielot et al., 2011]. Moreover, there is evidence which suggests that unfamiliar modalities are more likely to cause interruptions than modalities the user is constantly exposed to [Arroyo et al., 2002; Warnock et al., 2011]. This effect can be explained by the fact that more salient stimuli are more likely to cause involuntary attentional shifts (see Section 2.2.2). Thus, it is possible that the novelty of tactual feedback (as well as other exotic types of feedback) might outweigh any cognitive advantages, and cause the feedback to be more disruptive than its visual or auditory counterparts. Yet, training could help to reduce such effects [Cades et al., 2006].

Another disadvantage of tactual feedback is its reduced bandwidth. In particular, the spatial resolution of our tactile perception is mostly limited by the distribution of receptive fields on the skin. The finger, which is considered one of the most receptive part of the body [Weinstein, 1968], has receptive fields measuring upwards of 2 mm in diameter [Johansson, 1978]. Regarding the temporal resolution, studies show that humans are able to differentiate between tactile stimuli 2 - 40 ms apart, depending on location (Kenshalo 1978 in [Loomis, 1981]). However, these values are upper bounds which are very hard to reach due to limited attention and processing resources [Bliss et al., 1966; Hill and Bliss, 1968; Loomis, 1981]. More realistic thresholds are provided by Tan et al. [1999], who found that humans can optimally perceive 2-3 events per second using the haptic channel. In terms of minimal exposure time, Kaaresoja and Linjama [2005] suggest a minimum duration of 50 ms for reliable perception of vibrations.

There have been attempts of "improving" the bandwidth of tactual perception by defining meaning-infused patterns of vibrations, or placing the actuators at strategic locations on the body. Brewster and Brown [2004] provide a theoretical framework for designing tactile-interface. They introduce the concept of tactons, i.e. "structured, abstract messages that can be used to communicate complex concepts to users non-visually." A tacton is defined by seven parameters: frequency, amplitude, waveform, duration, rhythm, location on the human body and spatio-temporal pattern. The framework has been used in PocketNavigator [Pielot et al., 2010], where the authors encoded direction and distance to target by manipulating vibration rhythm and duration. The body location parameter has been successfully used by van der Linden et al. [2011b]. They placed actuators on the body parts which required the user's attention. Moreover, for each body part, two actuators were used and positioned opposing one another. This allowed them to also encode the direction in which the body part needed to be moved.

Overall, tactual feedback presents itself as a promising alternative to the more conventional and heavily used visual and auditory channels. In particular, tactual feedback scores highly relative to the second requirement of social augmentation (see Section 4.1) since it is less likely to cause cognitive interferences with the highly visual and auditory primary task of a social augmentation scenario. Yet, due to stimulus novelty, it is expected that this effect becomes apparent only after a longer exposure. However, the reduced bandwidth raises concerns regarding the ability of the feedback to deliver enough information to facilitate a behavioural change and thus satisfy R3. Relative to R6 (feedback protects privacy) and R1 (feedback can be perceived), tactual feedback shows promise since tactile stimulation is highly directional (only one person can perceive them[2]) and our skin is very perceptive of

[2]Here it is necessary to state that current vibro-actuators available on the market produce a fairly noticeable sound when turned on. Yet, this is a shortcoming of the devices and not of the modality.

Figure 6.4: Myo armband on user's forearm.

vibrations, regardless of the situation.

Vibrotactile Feedback during Social Interactions

To get a better image of the effectiveness of tactual feedback in a social scenario, we conducted a small user study [Damian and André, 2016]. 10 participants, aged between 21 and 33 (mean 27.4), 8 males and 2 females, tried out a simple tactual feedback system. The aim of the study was to measure how well the participants were able to detect different types of vibrotactile feedback while being engaged in a conversation.

The system itself consisted of a Myo armband[3]strapped to the user's forearm. While the Myo's main selling point is the EMG sensor, for the purpose of this study we only used its vibro-actuator. Using two parameters, intensity (value ranging from 1 to 250) and duration (in milliseconds), the Myo can be programmed to deliver various types of vibrations. To maximize the haptic sensation, the Myo was positioned with the LED-element (which also houses the actuator) on the underside of the forearm (see Figure 6.4). For the study, we implemented a total of 15 vibration patterns as detailed in Table 6.1. For the vibration scales, the starting values have been chosen following various pre-tests which showed that the Myo (running firmware 1.2.955) cannot produce vibrations shorter than 10 ms and of an intensity smaller than 60.

After receiving a short demonstration of the first five vibration patterns (P1 - P5), the participant engaged in a casual conversation with the experimenter for approximately 3 minutes. During this time, the Myo would vibrate 6 times, once every 30 seconds using a pattern chosen at random from the first five patterns in Table 6.1 (P1 - P5). The participants were instructed to mark the vibration occurrence and pattern in a questionnaire as soon as they felt it, and then continue with the conversation.

In the second part of the study, after the conversation had finished, the experimenter played back the two vibration scales (SD and SI). After each scale, the participant was asked how many vibrations they felt. During this part of the experiment, participants wore noise-cancelling headphones as the Myo's vibration-actuator also produces a noticeable sound when turned on, which can be more perceivable than the vibration itself.

[3]`http://www.myo.com`

ID	Duration	Intensity	Description
P1	500	150	Single short vibration
P2	500,500,500	150,0,150	Double vibration with 500 ms break
P3	1000	150	Single long vibration
P4	500	250	Single strong short vibration
P5	500,500,500	100,250,100	“wave” pattern
SD	10, 20, 30, 50, 100	150	Vibration duration scale
SI	500	60, 70, 80, 90, 100	Vibration intensity scale

Table 6.1: Vibration patterns. Duration is in milliseconds and intensity is a value from 1 to 250.

	P1	P2	P3	P4	P5
P1	12				
P2		12			
P3		1	10		
P4	5		1	6	
P5	1	2			9

Table 6.2: Confusion matrix showing recognitions of vibrotactile feedback events. Each event occurred 12 times during the study.

Upon analysis of the results, we noticed that the participants were able to recognize the patterns 81.6% of the time. The confusion matrix (Table 6.2) shows that they had a hard time recognizing vibration intensity, often times confusing P4 with P1. Despite the relatively high recognition rate, during the study most of the participants found the task of identifying the vibration pattern difficult, with P10 stating “this is harder than I thought.”

When asked whether the vibrotactile feedback disturbed the conversation (R2 in Section 4.1), most users said no. P3 stated “It may actually have been too subtle as I was very involved in the conversation.” P10 was more analytical, comparing it to visual feedback: “Since the vibrations use a different channel ... I was using my eyes to look at you ... I didn’t find [the feedback] disturbing.” A more thorough analysis of this effect will be provided in Section 9.

Regarding the vibration scales, the participants felt on average the last 2.2 vibrations from the duration scale, and the last 4 from the intensity scale. This translates to an approximated perception limen for the forearm of just under 50 ms for duration and 70 points for intensity. These findings are in line with the study by Kaaresoja and Linjama [2005] suggesting that the position of the Myo does not have a strong impact on the threshold. However, both results differ from the 2-40 ms distal threshold mentioned by Kenshalo (referenced in [Loomis, 1981]). One possible explanation is that the minimum temporal distance between two stimuli does not correlate strongly with the minimum duration of a stimulus. The difference might also be caused by the fact that the vibration-actuators in both the Myo and the mobile phone used in [Kaaresoja and Linjama, 2005] are not in direct contact with the user’s skin.

6.1.4 Thermal Feedback

Thermal feedback represents the transfer of information to the user in response to an action using thermal stimulation (i.e. cooling or heating) of the user's skin. Similar to our tactual sensing, thermal perception also involves cutaneous stimulation yet the subjective sensation one feels from thermal stimulation is very different from a tactual stimulation. Regardless of how we categorize it, thermal feedback is a promising channel for communicating information to the user and has seen a surprising amount of interest from the research community over the past 10 years.

Every person has a neutral zone for skin temperature which measures 6–8 °C and is situated between 28°C and 40°C [Jones and Berris, 2002; Stevens, 1991]. Temperature changes within the neutral zone of a person are less likely to be perceived [Jones and Berris, 2002], but not impossible [Halvey et al., 2013]. Within this neutral zone, adaptation can occur (a new temperature feels like neutral after a period) if the user is exposed to a constant stimulus [Jones and Berris, 2002]. Outside the neutral zone, adaptation does not occur any more regardless of the stimulation intensity or duration, causing the user to feel constant warmth or coldness [Kenshalo et al., 1968].

Empirical evidence also suggests we are better at detecting fast changes of temperature. More specifically, the minimum amount of temperature change which is perceivable, called just noticeable differences (JND), is linked to the rate of change (ROC) in temperature. In most cases, we are able to perceive smaller changes if the ROC increases. This rule only holds up to a ROC of 3°C/s, after which the JND increases again [Claus et al., 1987; Harrison and Davis, 1999; Pertovaara and Kojo, 1985]. The JND also varies with the location of the stimulation and the type of skin it is applied to. Generally, hairy skin is more sensitive (lower JNDs) than hairless (glabrous) skin as the latter tends to be thicker [Harrison and Davis, 1999; Pertovaara and Kojo, 1985]. The sensitivity decreases towards the extremities of the body, with the head and trunk being the most thermal sensitive locations [Claus et al., 1987; Hagander et al., 2000]. Moreover, temperature sensation is also known to be influenced by clothing [Halvey et al., 2011], mobility [Wilson et al., 2011] and environmental factors [Halvey et al., 2012b].

Recent user studies put these theories to the test in an HCI scenario and yielded some general design guidelines for working with thermal feedback. Halvey et al. [2013] state that "for a change to be perceived, a minimum [temperature change] of 3°C should be used", although in their study, they found that even small changes (1–2°C) can be perceived, albeit less accurately.

Other studies have shown that temperature can have a large effect on our social interactions, being able to induce social proximity and affect our choice of words [IJzerman and Semin, 2009]. This result is backed by Williams and Bargh [2008], who found that users linked the experience of physical warmth to interpersonal warmth. Other researchers found that warm stimuli were perceived as more arousing [Salminen et al., 2011] and more pleasant [Halvey et al., 2012a] than cold stimuli.

Several researchers have also looked at how thermal feedback can be used to deliver information to the user. Wilson et al. [2015] explored the inherent meaning of thermal stimuli during daily interactions. Overall they found that warmth communicates recent activity, physical presence, higher content use and positive experiences, and cold feedback communicates the opposite. This study is in line with other results which found that warm stimuli were generally mapped to positive aspects (physical attraction, enjoyment, gratitude, good quality) whereas cold stimuli were mapped to negative aspects (nervousness, sadness,

pain, strangeness, bad quality) [Lee and Lim, 2010, 2012; Suhonen et al., 2012].

In the HCI domain, the most common way of delivering thermal feedback is by using Peltier elements which are able to generate or remove heat when current flows through a junction of two conductors. Researchers have used Peltier elements to deliver thermal feedback to the palm of the hand [Halvey et al., 2013], forearm [Lee and Lim, 2010; Wilson et al., 2015], upper arm [Wilson et al., 2015] or torso [Gooch and Watts, 2010] in both static in-front-of-pc [Halvey et al., 2013] and mobile scenarios [Lee and Lim, 2010]. Interpersonal communication seems to be the most popular type of application involving thermal feedback. Suhonen et al. [2012] used the haptic and thermal modality for enhancing experiences in remote communication systems. They conclude that thermal stimuli lend themselves well to communicating valence. Researchers have also used to thermal feedback to exchange "thermal massages" [Lee and Lim, 2010], give "thermal hugs" [Gooch and Watts, 2010] or enhance the affective bond of a user with their teddy bear [Willemse et al., 2015].

To sum up, the thermal modality has seen a growing amount of support in the last decade which revealed that it is a reasonable channel for delivering information to the user. In the social augmentation scenario, thermal feedback represents a valuable alternative to the overused auditory and visual channels, since it targets the cutaneous sense which is mostly idle during a social interaction. Thus, thermal feedback promises disruption free information delivery and thus places the modality high relative to the second requirement of social augmentation (see Section 4.1). However, as discussed before, the novelty of the modality could (at least initially) counteract this effect and lead to an increased disruption of the primary task [Arroyo et al., 2002]. Moreover, the inherently reduced bandwidth of the modality combined with a high susceptibility to environmental factors raise concerns regarding the ability of thermal feedback to deliver sufficient information to facilitate a behavioural change (R3) or to be perceived at all (R1).

6.1.5 Olfactory Feedback

The sense of smell is profoundly intertwined with our lives, albeit in a less obvious fashion than some of our other senses. We use smell to continuously gather information about our surroundings. Smell tells us if our food is good, warns us from harmful gases and even helps us find the right partner [Spehr et al., 2006]. We detect smell either through the nose (the orthonasal pathway) or through the back of the mouth (retronasal pathway). Using these pathways, air is directed to a pair of large cavities on top of which the actual sensing tissue resides (olfactory epithelium). How exactly scent is detected is still not fully understood [Amoore et al., 1964; Turin, 1996]. This combined with the high variability of olfaction between individuals [Gibbons, 1986; Stevens et al., 1988] and across time [Berglund et al., 1971; Lawless, 1997], makes the classification of scents an especially challenging and yet unsolved task [Kaye, 2001].

Thus, unlike with vision where we know that most colours can be expressed as combinations of red, green and blue, we have no knowledge of how to "generate" scents. Most current olfactory applications pulverize essential oils or liquid compounds into the air. Such approaches are limited in complexity (one liquid compound can only generate one scent) but also in their delivery mechanism. Range, decay rate and direction are difficult to control due to air dynamics, as highlighted by the following reaction to the AromaRama, a smell-enriched cinema: "At one point, the audience distinctly smells grass in the middle of the Gobi desert"(Time Magazine, quoted in [Kaye, 2001]). Furthermore, the olfaction variability

mentioned in the previous paragraph means that one person may smell something different than another person. I noticed this effect recently on a wine tasting tour in western France where, as a game, the tour guide had us smell and guess different wine scents. I was stunned to see that in the majority of cases, every single person in the room thought to be smelling a different scent. Needless to say, designing a complex user interface under these conditions is borderline impossible.

Despite challenges, olfactory feedback has been used in various scenarios. Japanese incense clocks burned steadily throughout the day and different incense tablets marked the hours allowing monks to tell the time by simply sniffing the air [Bendini, 1964]. In recent years, a number of attempts for computerized olfactory user interfaces have been made. Joseph Kaye's [Kaye, 2001] wittily named application "Dollars and Scents", uses two scents (lemon and peppermint) to give feedback to users regarding changes in a particular stock. While an initial informal test had a positive response, no formal evaluation has been conducted. Furthermore, Kaye acknowledges that problems with scent delivery and perception makes it difficult to engineer a more complex interface. Bodnar et al. [2004] presented a concept for an olfactory display to deliver information to users which were engaged in another task. An evaluation of their system showed that while olfactory feedback was less obtrusive than audio or visual feedback, it was also much less effective. The poor performance was also reported by Warnock et al. [2011]. When comparing between visual, auditory, tactile and olfactory notifications, they found that olfactory notifications had "the longest response time" and "the lowest number of correct responses." Interestingly, unlike Bodnar, Warnock also found evidence that olfactory message were more disruptive towards the primary task, causing a "decrease in player speed compared to visual or audio notifications." This increase in disruptiveness was also reported by Arroyo et al. [2002], which attributed it to the overall novelty of the stimulus. Yet, it is plausible that repeated exposure to the stimulus could reduce this effect.

There has also been work done in the automotive domain which use smells to increase a user's alertness and combat drowsiness. Yoshida et al. [2011] developed a system which releases various scents (peppermint, rosemary, eucalyptus and lemon) whenever the driver exhibits drowsiness symptoms. A study by Raudenbush et al. [2009] found that a periodic administration of cinnamon and peppermint scents "led to increased ratings of alertness, decreased temporal demand, and decreased frustration" during a simulated driving test. Currently, there are some commercially available hardware solutions for delivering olfactory feedback. However, a recent review showed that these vary greatly in terms of delivery speed, volume and distance [Dmitrenko et al., 2016].

Overall, despite the obvious advantage that olfaction is less used in primary tasks and therefore olfactory feedback is less likely to interfere with them in a social augmentation context (R2 — see Section 4.1), the inherent problems associated with manipulating odours, and the poor performance of olfactory displays raise concerns regarding the ability of olfactory feedback to fulfil R1 (feedback is perceivable) and R3 (feedback facilitates change in behaviour).

6.1.6 Gustatory Feedback

From all the human senses, taste is the one which, at first glance, appears to be the least compatible with HCI. The main reason for this is the small and enclosed area which holds the receptors. This makes the process of gustatory stimulation highly invasive, as it involves

direct contact with the interior of the user's mouth.

Unlike the classification difficulties of olfactory perception, researchers mostly agree that there are five basic taste categories: sweet, sour, salty, bitter and umami. The only exception is pungency (i.e. hotness), which is also considered a type of taste in some Asian countries. However, since the hotness (e.g. of a chilli) is not actually perceived by the taste receptors but by a direct stimulation of the nerves, most researchers do not consider it a taste. Furthermore, in 2015 researchers proposed adding "fatty taste" as a new basic taste category named oleogustus [Running et al., 2015]. These basic tastes are seldom experienced individually but rather in various combinations to form the complex sensations we so enjoy. Keast and Breslin [2003] found that tastes combine in very interesting ways, with some having the ability to override others or, in the case of sweet and sour, can even cancel each other out. Another important characteristic of taste is its diachronicity, or temporal evolution. Researchers have found that different tastes evolve differently over time. For example, salty tastes are more dynamic than sweet tastes [Nakamura et al., 2012; Pineau et al., 2009], bitter and umami persist longer, and sour has the shortest duration [Obrist et al., 2014] . However, not all persons perceive taste the same way which has led to the introduction of a "taster status" [Bartoshuk, 2000]. According to this, persons can be categorized according to their taste sensitivity into: supertasters (25% of the population), medium tasters (50% of the population) and non-tasters (25% of the population) [Chen and Engelen, 2012].

Surprisingly, there have already been some attempts at using gustation in the HCI domain. Murer and colleagues presented the "LOLLio" [Murer et al., 2013], a device which features accelerometer-based input and chemically-induced gustatory output, all in one mobile enclosure. The device has been successfully used in a children's game [Moser and Tscheligi, 2013] making it the first real application to employ gustatory feedback. A different approach to gustatory stimulation is to use electrical and thermal stimulation to generate sour, bitter and salty tastes [Ranasinghe et al., 2012]. This has the benefit of not relying on consumable substances. However, it is more limited in the range of taste experiences it can generate.

Despite the sensorial prowess of the human tongue, taste is generally considered to be a multi-sensory experience also involving smell (picked up through the retronasal pathway), mechanoreception (food texture), thermoception (food temperature) and even sound. Considering this, many researchers also experimented with taste in a multi-modal context. Iwata et al. [2004] presented the "Food Simulator", a device able to simulate both gustatory, haptic and auditory properties of various types of food. Ranasinghe et al. [2015] combined electrical and thermal gustatory stimulation with chemically based scent emitters to generate "digital flavours."

Similar to tactual and olfactory feedback, the largest benefit of gustatory perception is its highly specialized area of use. Our gustatory system is idle any time we are not eating or drinking. This yields a large theoretical potential for social augmentation systems as it promises an interference-free input channel. This places gustatory feedback high relative the second requirement of augmenting social interactions, which states that feedback should not impact the attentional level dedicated to the primary task (Section 4.1). Yet, similar to the other exotic feedback modalities, a novelty effect is to be expected which would increase the cost of the interruptions. Moreover, the large variations in gustatory sensation across the population as well as the ambiguity regarding the interaction between different tastes places gustatory feedback at odds with R1 (feedback can be perceived). The intrusiveness associated with gustatory stimulation is also likely to break social conventions (R5).

6.1.7 Multimodal Feedback

While up until this point, this chapter mostly listed unimodal feedback examples, it did this in an effort to make the proposed classification clearer and easier for the reader to understand. Multimodal feedback, i.e. feedback spanning multiple sensory input channels, is not only possible, but highly encouraged especially in situations where cognitive overload might render some perception channels unresponsive. In such situations, transmitting information on multiple channels has the benefit of not relying on the availability of a single type of cognitive resources for successful perception, but would allow the user's perception mechanism to freely "choose" the channel from which to extract the information.

In everyday life we are used to multimodal sensing. Studies have found that hearing somebody talk is less effective than seeing and hearing at the same time [Campbell et al., 1998]. Modern attention models suggest that different modalities use different cognitive resources, and that providing information using one modality is less efficient than distributing the same amount of information across multiple modalities [Burke et al., 2006; Wickens, 2002]. This effect has been also reported in studies where users favoured multimodal feedback as task complexity increases [Oviatt et al., 2004]. This suggests that distributing information across multiple modalities allows for a more balanced use of mental resources and might prevent cognitive overload [Sigrist et al., 2012].

There already is a large body of work which focuses on various research topics pertinent to multimodal interaction. Furthermore, applications which deliver multimodal feedback by targeting multiple senses are getting ever more popular. The most common type of multimodal feedback is achieved through the combination of the visual and the auditory modality. William Gaver [Gaver, 1989] was one of the first to notice the potential of multimodal feedback for HCI:

> "Sound plays an integral role in our everyday encounters with the world, one that is complementary with vision. It is in understanding this role that the most compelling reasons for using sound [in HCI] become clear."
>
> (William W. Gaver [Gaver, 1989])

Since then, most operating systems have used auditory icons in parallel with visual feedback, for example when opening on a file or navigating through the file explorer. Empirical evidence has also been reported which backs up Gaver's statement. Both in flight and driving simulators, audio-visual feedback has been found to increase performance [Bronkhorst et al., 1996; Liu, 2001]. Since in such tasks the visual channel transports the most information, it is expected that moving some of this load to another modality would be beneficial [Wickens, 2002]. Studies have also shown that especially for older adults, multimodal feedback yields an increase in performance when interacting with mobile devices [Lee et al., 2009]. The aforementioned "Food Simulator" [Iwata et al., 2004] also uses three different modalities to create the impression of taste. Here, the authors themselves state that unimodal stimulation is not sufficient for a believable simulation of taste. Multimodal feedback can also improve the ability of systems to convey emotions, enabling an overall larger "affective range" when compared to unimodal solutions [Wilson and Brewster, 2017].

Overall, multimodal feedback can reduce cognitive load and improve user performance for a given task, making it an ideal candidate for our social augmentation scenario. If we apply Coutaz et al.'s CARE properties [Coutaz et al., 1995], we end up with the following configurations for multimodal feedback (also illustrated in Figure 6.5):

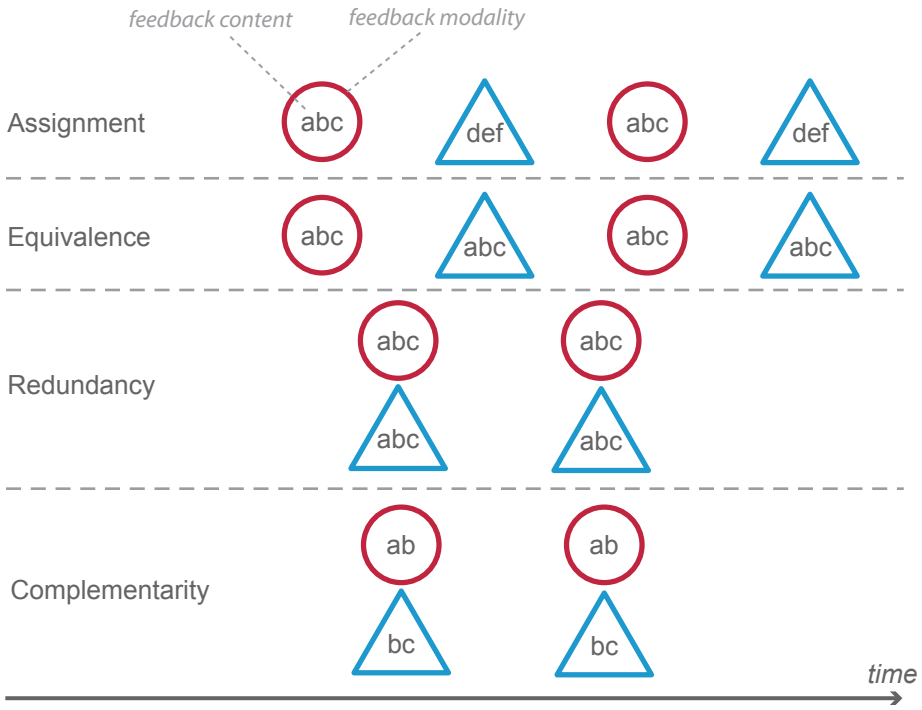

Figure 6.5: The four multimodal feedback configurations for social augmentation.

- *Assignment*: A single modality is used to deliver feedback regarding one specific behaviour. Assignment is the standard unimodal feedback configuration found in countless systems: A microwave oven beeps when the program is finished or a warning light flashes when something is amiss. In a social augmentation context, an auditory beep could be used to inform the user they are smiling too much.
- *Equivalence*: Two or more modalities are used alternately to deliver feedback regarding the same behaviour (but not at the same time). For example, both sound and visual feedback are used alternately to inform the user of an event. This configuration makes most sense in adaptive interfaces [Arroyo et al., 2002; Arroyo and Selker, 2003] which dynamically switch between modalities in an effort to search for the most effective one.
- *Redundancy*: Two or more modalities are used simultaneously to deliver feedback regarding one specific behaviour, and the feedback events are carrying the same information. The most common examples of redundancy can be found in modern graphical user interfaces: Performing an action triggers both a visual feedback (e.g. animation) as well as an informationally redundant auditory one (e.g. beep). Similarly, a mobile phone will ring and vibrate at the same time to inform the user of an incoming call. In the context of social augmentation, a sound paired with a short vibration on the forearm could inform the user they are smiling too much. In this case, two feedback events are delivered simultaneously to form one multimodal feedback event.
- *Complementarity*: Two or more modalities are used simultaneously to deliver feedback regarding one specific behaviour, and the feedback events are not carrying the same information but are complementing each other. A good example of complementary feedback is Iwata's FoodSimulator [Iwata et al., 2004], which uses tactual and auditory feedback to deliver information on the texture of the food item, and gustatory feedback

for communicating taste information. In a social augmentation scenario, a vibration on the forearm could inform the user that the smiling behaviour is inappropriate whiles a simultaneous visual feedback gives more details on how to correct it (e.g. “smile less”). Here, two informationally distinct feedback events are delivered simultaneously to form one multimodal feedback event.

For the social augmentation scenario, the redundancy configuration poses two main advantages. First, when delivering feedback across multiple channels, the individual perceptual weaknesses of each modality become less important. For example, the poor reaction time associated with olfactory feedback is less of an issue if it is accompanied by visual feedback. Secondly, multiple channels providing perceptual redundancy means that if the user misses the feedback event on one channel, they may still perceive it on a different channel. This is particularly useful in stressful situations where the user’s cognitive system is already overloaded with processing the primary task, making information loss more likely. This effect is directly related to the resource-based attention models introduced in Section 2.2. These suggest that performance degradation is more likely to happen for tasks sharing the same cognitive processing structures. Thus, for a behavioural feedback loop, the best case is to provide feedback on a perceptual channel which is not used by the primary task. However, since directly measuring the impact of a task on the cognitive system of the user is not an option (at least not without greatly impairing the user’s ability to participate in live social interactions), the only other way to find an “empty” channel is to “try” all of them out using multimodal feedback. This gives the behavioural feedback loop the best chance of avoiding a potential cognitive bottleneck caused by the primary task. Thus, redundancy scores well relative to the first requirement of augmenting social interactions (feedback is perceivable — see Section 4.1).

Complementarity inherits the advantages of the redundancy configuration and adds one more. With complementary feedback, each feedback event can be tailored to the advantages of its modality. For instance, the large bandwidth of the visual channel can be used provide detailed information on the behaviour analysis while the natural affective mapping of thermal events can encode the appropriateness of the behaviour in the current situation. Thus, more information can be transmitted using complementary feedback, increasing the likelihood that the user will know how to change their behaviour (R3). However, the risk arises that if the individual feedback events are very dissimilar and one feedback event is not perceived, the user will not be able to extract the meaning of the multimodal feedback event. This issue can be addressed by adding some redundancy to the configuration so that each modality carries enough meaning to make it understandable on its own, and only the details are complementary between the modalities.

The assignment configuration is also interesting as it strengthens the user’s mapping of the feedback events to the underlying behaviours. For instance, always delivering feedback concerning the verbal behaviour using the auditory modality and feedback related to body behaviour using the tactual modality helps strengthen the mental bonds between perception, reflection and action. Thus, it may have a positive effect on the automation of the behavioural feedback loop which would translate to a diminished impact of the social augmentation on the primary task (R2) and an increased likelihood of reacting to feedback events (R3). Furthermore, some studies have shown that multimodal feedback may also lead to an increased cognitive load in certain high workload scenarios [Burke et al., 2006]. In such situations, an unimodal configuration might be preferred.

Finally, as discussed above, equivalence makes most sense for feedback adaptation. However, switching back and forth between modalities can be damaging to the social augmentation since it may disrupt the mental mapping the user has formed between the feedback and the cause of the feedback. This might interfere with the feedback's ability to elicit the intended behavioural change from the user (R3).

6.2 Prominence

The prominence characteristic defines how perceptually striking a feedback event is and thus how attentional demanding it can be. Attentional demand was introduced by Wendy Ju in her framework for implicit human-computer interactions [Ju, 2015] as means to differentiate between background and foreground interactions. According to her, attentional demand represents "the attention demanded of the user by the computer system" and "can be manipulated by adjusting the perceptual prominence of objects." Whereas the timing of the feedback is able to control the cost and the frequency of interruptions (see Section 4.3.3), the prominence impacts the probability that a feedback event will produce an interruption of the main task. This is because highly prominent (salient) stimuli have a larger likelihood of diverting the attentional focus away from the primary task (see Section 2.2.2).

Although prominence is modality dependant, in most cases, some form of stimulus "intensity" is available:

- *Visual feedback*. Factors which affect the perceived intensity of visual stimuli are the brightness of the display, the contrast between foreground and background, the size and location of the stimulus, and use of specific colours [Lupton, 2004]. Moreover, using animations (i.e. moving images) is also known to be more prominent than static images.
- *Auditory feedback*. For the auditory channel, the sound pressure level and pitch of the sound have been found to correlate with the perceived loudness of a sound [Olson, 1972; Stevens, 1955].
- The perceived intensity of *vibrotactile feedback* can be directly controlled by the amplitude of vibration [Verrillo et al., 1969] and area of stimulation [Cholewiak, 1979].
- For *thermal feedback*, the difference between the temperature of the stimulus and the skin temperature as well as the temperature rate of change correlates to the perceived intensity of the stimulus [Jones and Berris, 2002]. Similarly to tactual feedback, the perception of temperature also varies with stimulus location [Claus et al., 1987; Hagander et al., 2000].
- *Olfactory feedback*. Although the intensity of an odour is formally defined as a function of chemical concentration [Jiang et al., 2006], subjective intensity is also influenced by the hedonic tone (or pleasantness) of the stimulus. That is, unpleasant odours tend to be considered more intense.
- *Gustatory feedback*. Similarly to olfactory stimuli, perceived taste intensity is also governed by chemical concentration [Smith, 1971]. However, each taste has its own scale and perceived intensity [Pfaffmann, 1980]. Overall, bitter tastes are considered more unpleasant and thus can elicit a stronger reaction from the user. The same holds also for the (non-taste) sensation of spiciness, which can also quickly become unpleasant.

Besides this "intensity" factor, there is also evidence that combining multiple modalities leads to an increase in prominence. Politis et al. [2013] found that multimodal feedback is

Figure 6.6: The prominence dimension relative to the first two requirements of social augmentation.

associated with a greater sense of urgency than unimodal feedback.

Going over the related work, one can notice that systems which rely on realtime visual feedback often employ low prominence feedback in an effort to be unobtrusive and subtle [Boyd et al., 2016; Damian et al., 2015b; Tanveer et al., 2015]. However, in some cases, such as Schneider's PresentationTrainer [Schneider et al., 2015], high prominence feedback is explicitly triggered whenever "severe" behavioural mistakes are detected. For vibrotactile feedback, determining the intensity is generally difficult since most consumer-ready actuators either do not support manual amplitude control or provide discrete pre-defined amplitude profiles (low, medium, high) for which no documentation is available. In the case of systems using thermal feedback [Gooch and Watts, 2010; Suhonen et al., 2012; Wilson et al., 2015], their primary aim is to provide comfortable feedback. Thus, the feedback can be considered low prominence, with temperatures varying between 20 and 39° Celsius. The same philosophy was used for the gustatory and olfactory feedback systems, with most of them opting for low concentrations of pleasant tastes [Murer et al., 2013] and odours [Kaye, 2001].

For social augmentation, a low prominence level results in subtle, less distracting but also easier to miss feedback. A high prominence increases the chance that the feedback will be noticed by the participant but might also distract them from the primary task. Thus, choosing the right prominence represents an act of balance between the first (feedback is perceivable) and second requirement (feedback is not disruptive) of augmenting social interactions (see Section 4.1). This relation between prominence and the social augmentation requirements is illustrated in Figure 6.6.

The following sections will now look at two concrete approaches towards managing feedback prominence: ambient information systems and subliminal feedback.

6.2.1 Ambient Information Systems

One specific type of systems which actively attempt to minimize feedback prominence is ambient information systems. The term ambient information system has been coined by Pousman and Stasko [2006] and refers to systems which deliver information "in a way that is not distracting, but is aesthetically pleasing." The main characteristics of such systems are that they target the periphery of the user's attention [Heiner et al., 1999] and blend into the physical environment [Hazlewood et al., 2008]. More precisely, Pousman and Stasko [2006] defined a total of five characteristics which define ambient information systems: (1) the delivered information is important but not critical, (2) the systems target the periphery of the user's attention (but can move to the focus and back again), (3) focus on (tangible) representations in the environment, (4) are subtle and not distracting, and (5) are aesthetically pleasing and environmentally appropriate.

One typical example of an ambient information system is a string that wiggles according

to network traffic [Weiser and Brown, 1995]. Light is also a common element in such systems. Chang et al. introduced Lumitouch [Chang et al., 2001] — a pair of picture frames where one subtly lights up when the other one is touched. The system is designed to have a low prominence level and not draw the user's attention, but rather provide information if and when the user looks at the frame. Similarly, Ambient Timer [Müller et al., 2013] uses colour in the user's peripheral vision to provide information on upcoming appointments. By gradually changing the colour, the system provides a "non-distracting way of monitoring the remaining time [until the next appointment]."

Ambient information displays are commonly associated with a decrease in attentional footprint. Yet, most ambient information displays lack formal evaluations [Mankoff et al., 2003] as the authors mostly focus on technology and design. There are some exceptions. For instance, Müller et al. [2014] found that ambient feedback caused less cognitive workload when compared to on-screen visualizations. Similarly, Occhialini et al. [2011] found that ambient displays can be read at a glance "without distracting or annoying the users."

As with all types of low prominence feedback, using ambient feedback in a social augmentation scenario can be risky since it might get ignored by the user (violating R1 — see Section 4.1). The next section will discuss how low prominence feedback which is not perceived consciously can still impact the behaviour of the user without them being aware of it.

6.2.2 Subliminal Feedback

According to the concept of subliminal psychodynamic activation (see Section 2.2.3), in case of a subliminal perception of feedback, the stimulus is perceived, processed and interpreted outside of the user's awareness. Thus, the user is unaware of the presence and the meaning of the feedback message, but the stimulus may still influence the user's mental state and actions.

So far, only few attempts have been made in the HCI context to use subliminal feedback. In 1991, Wallace et al. [1991] studied the effect of subliminal visual stimuli on learning to operate a text editor on a desktop computer. They did find some differences between users receiving subliminal help and users which do not in terms of duration before explicitly requesting help. However, these differences were not significant. Twelve years later, a research group from MIT introduced the "Memory Glasses" [DeVaul et al., 2003], which provide memory support using subliminal visual cues presented on a head-mounted display (HMD). They conducted a user study in which the participants would have to recall the name associated to a given face. To help the users remember, the correct name was shown subliminally (5 ms duration) on the HMD for each face. They found that the subliminal cues did improve the users' ability to associate names to faces. Although the reported differences are significant, they do not seem to be substantive[4]. Barral et al. [2014] made use of a virtual environment to perform subliminal priming. More specifically, they asked users to choose one of two food items in a virtual fridge. Before starting, the users were shown a visual subliminal cue (33 ms in duration) representing one of the food items. The study revealed that the subliminal cues were able to influence user behaviour but only for "fast" users, i.e. instances where the users made a choice in under a second. And even then, the accuracy in choosing the primed food item

[4]Without subliminal cues, the users managed to recognize on average 1.62 faces out of 21. With subliminal cues, this value increased by 0.76. Assuming that the reported "effect σ^2" represents the variance of the population, the effect size can be calculated to $r = 0.105$. This value is quite small and suggests that the results are indeed not substantive.

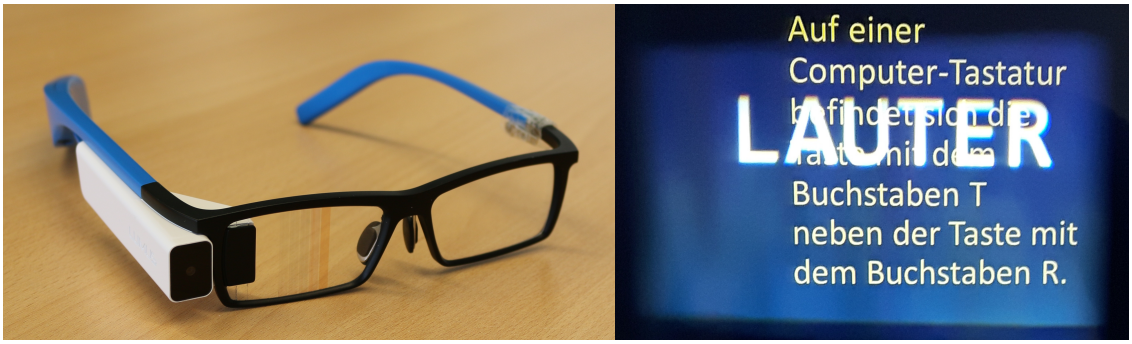

Figure 6.7: The Lumus DK-40 HMD (left) and the view through the HMD showing the stimulus instructing the user to speak louder (right).

improved only by 13% to a total of 66%. Thus, the authors conclude that "in more natural conditions, it is expected that the cues would mostly go unobserved."

Overall, despite the theoretical beneficial effects of subliminal feedback, current implementations show that in practice, subliminal feedback is far from being able to accurately influence behaviour. This raises concerns regarding the ability of subliminal feedback to reliably satisfy the first and third requirements of augmenting social interactions — feedback can be perceived and is appropriate for facilitating a behavioural change (see Section 4.1). These concerns are put to the test in the remainder of this section.

Using Subliminal Feedback in a Behavioural Feedback Loop

To get an impression of how well subliminal feedback would fit in a social augmentation scenario, we [Bottari, 2017] integrated subliminal feedback in a behavioural feedback loop and tested its effectiveness with the help of a user study.

Study Design

A total of 20 participants, 13 male and seven female, with a mean age of 27.35 have been recruited for the study. Each participant was tasked with reading out a text from a remote monitor. During the reading task, the participants were given visual feedback on their speech loudness using a Lumus DK-40 HMD (depicted in Figure 6.7 left). To eliminate variability, the feedback was always the same: It instructed the participant to speak louder using a textual representation (see Figure 6.7 right), regardless of how loud (or soft) the participant actually spoke.

The study tested four conditions: two subliminal and two supraliminal conditions. In the first subliminal condition (*sub*), a stimulus consisting of the text "LAUTER" (German for "louder") was delivered subliminally to the user. To achieve subliminal perception, the masking paradigm [Marcel, 1983] was employed. More precisely, the stimulus was presented briefly on the screen of the HMD, and a mask was displayed before and after the stimulus. The stimulus had an exposure of approximately 18 ms[5] whereas each mask was displayed for 200 ms. The second subliminal condition (*sub′*) acted as a control. Here, a different stimulus that consisted of a nonsensical character sequence was displayed for the same amount of time.

[5]The stimulus was rendered for exactly one frame, which theoretically should result in a display duration of 16.66 ms (the HMD had a 60 Hz refresh rate). However, measurements prior to the study using a high-speed camera yielded an average exposure of 18.07 ms over 50 repetitions. The difference can be explained by inaccuracies in the rendering pipeline on the HMD.

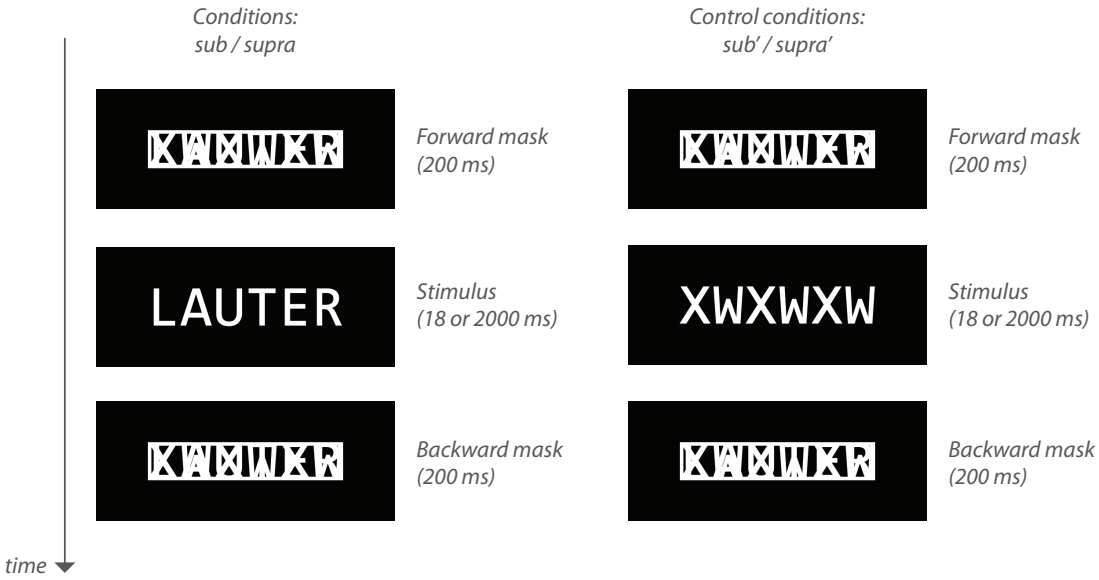

Figure 6.8: The design of the study with all four conditions.

For the supraliminal conditions (*supra* and *supra*′), the same stimuli were displayed but for two seconds each. For consistency, the masking was used in every condition, before and after the stimulus. Figure 6.8 illustrates the study design.

The conditions were carried out in a within-subjects fashion, and their order was randomized. In each condition, the stimulus was repeated five times, once every 30 seconds. Thus, a session lasted for approximately 10 minutes, during which each participant was exposed to a total of 20 stimuli.

Results and Discussion

During the experiment, the intensity of each participant's voice has been computed and recorded using the HMD's integrated microphone and an SSJ pipeline (see Section 7). Post-hoc data analysis showed that on average, the vocal intensity increased significantly after being exposed to the real supraliminal stimulus ($M_{supra} = 34.25$ dB) when compared to the control stimulus ($M_{supra'} = 33.34$ dB, $p < 0.01$). However, no such effect has been observed for the subliminal conditions ($M_{sub} = 33.02$, $M_{sub'} = 33.13$). Moreover, when comparing between the subliminal and supraliminal conditions, paired T-Tests showed significant differences between the *sub* and *supra* conditions ($p < 0.01$). The differences between the two control conditions (*sub*′ and *supra*′) were not significant.

The results suggest that the feedback was able to influence the participants' loudness, yet only in the supraliminal condition. Participants did not adapt their behaviour in response to the subliminal feedback. However, it is unclear whether this lack of a reaction was caused by a complete lack of perception of the subliminal stimulus, or just the inability of the subconscious to process the feedback and trigger a reaction.

Overall, the study confirmed the suspicions postulated before. Despite its theoretical marvel, subliminal feedback is ill suited for deployment in real HCI applications, especially in a social augmentation scenario. It might be that for subliminal behavioural feedback loops, different types of feedback might be better suited. The stimuli in the famous MIO

studies [Bornstein, 1990; Silverman and Weinberger, 1985; Weinberger, 1992] were given an emotional significance by invoking the participants' relationship with their mothers with the help of the text "Mommy and I are one." Moreover, fMRI studies suggest that emotional faces (e.g. fearful or angry) induce a stronger activation in the brain than neutral stimuli [Brooks et al., 2012]. Another approach would be to make explicit use of fully unconscious cognitive mechanisms such as those governing our social interactions. When we interact with other persons, we do so on multiple levels and we are unaware of most of these. One particular example of such a mechanism is mirroring. Mirroring, or the "chameleon effect", is defined as the "nonconscious mimicry of the postures, mannerism, facial expressions and other behaviours of one's interaction partners" [Chartrand and Bargh, 1999]. Empirical studies have found that humans employ mirroring with persons they respect and have a positive attitude towards [Hess and Fischer, 2013] and is critical for building rapport [Iacoboni, 2008]. Yet, instances of using social subliminal feedback in an HCI context are quite rare. Bailenson and Yee [2005] used head movement mirroring in an interaction scenario between a human and a virtual agent. They found that "mimicking agents were more persuasive and received more positive trait ratings than nonmimickers, despite participants' inability to explicitly detect the mimicry." However, these social or emotional stimuli have one thing in common: They carry little to no information. Thus, it is unlikely whether such approaches could be used to inform the user of their behaviour quality as part of a social interaction scenario.

6.3 Duration

Once the feedback has been delivered, its duration defines how long the feedback message is perceivable by the user. A short duration is specific of auditory (e.g. beeps) and tactile (e.g. vibration pulses) feedback whereas visual feedback is typically delivered with a longer duration (e.g. images, video).

Looking at related work we can see different implementations for visual feedback ranging from low duration timeout-based solutions to persistent always-on feedback. Both Tanveer et al. [2015] and Boyd et al. [2016] opted for low duration solutions, displaying visual feedback for up to three seconds. The largest advantage designers hope to get from this type of feedback is a reduced disruption effect due to limited exposure. However, there is no empirical data to support this claim.

On the other side of the spectrum, the increased exposure of long duration feedback gives the user more time to perceive, interpret and react. Thus, it minimizes the chance of the feedback being missed by the user. This is particularly likely to happen in stressful situations where the cognitive demand of the primary task is very high, leaving no spare cognitive resources for a secondary task. Moreover, psychological literature suggests that a longer exposure might actually reduce prominence (and thus its attentional demand) since the stimulus becomes familiar to the user [Johnston et al., 1990]. An example of long duration feedback can be found in Schneider's PresentationTrainer [Schneider et al., 2015]. Here, visual feedback is displayed continuously until the cause of the feedback is eliminated (i.e. the user corrected their behaviour). Similarly, van der Linden's MusicJacket [van der Linden et al., 2011b] delivers continuous vibrotactile feedback until the user corrects their posture. Section 8 will introduce a system where visual feedback is displayed persistently, allowing the users to always inspect the quality of their own behaviour.

Ofek et al. [2013] directly compared between short and long duration visual feedback.

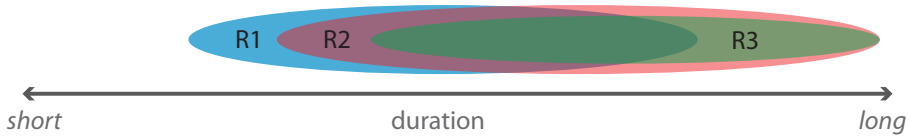

Figure 6.9: The duration dimension relative to the first three requirements of social augmentation.

They found that the long duration solution (message is displayed until user reacts) was more likely to trigger a reaction, had shorter response times and was less disruptive than the short duration feedback (message is displayed for three seconds). However, subjectively, the users preferred the short duration feedback.

To sum up, in a social augmentation context, longer durations are generally more appropriate, increasing the likelihood of perception (R1 — see Section 4.1) and of eliciting a reaction from the user (R3). Moreover, since there is evidence that an increased exposure is associated with a reduction in prominence, it also scores well relative to the second requirement. However, persistently displayed feedback may cause the feedback event to become familiar and lose its ability to involuntary capture our attention (see Section 2.2.2). In this case, the augmentation would rely solely on voluntary attentional shifts, placing it at odds with R1. Figure 6.9 illustrates these considerations.

6.4 Scope

The scope of the feedback defines whether the feedback simply informs the user of the current state of their behaviour (e.g. “you smile rarely”) or whether it represents and instruction to change to a different behavioural state (e.g. “smile more”). The first type, called appraisive feedback, implements the classical feedback loop paradigm: It presents objective information on the current state of the user’s behaviour and asks the user to decide if and what behavioural change is needed. A typical case for appraisive feedback is the radar speed sign example presented in Section 4.2. Here, the speed is measured and displayed to the driver, but no explicit instruction to slow down is provided (see Figure 6.10-a). The driver is expected to reach this conclusion by themself. Often, appraisive feedback is also enhanced with interpretations of the current behavioural state. In the speed sign example, these take the form of smileys which are either positive when the speed is below the speed limit or negative when it is above (see Figure 6.10-b).

The second type is called instructional feedback and closely follows the coaching metaphor as it aims to give clear instructions on how to adapt one’s behaviour to reach a more desirable state (see Figure 6.10-c). This clearness is also its main benefit, as it eliminates the need for the user to interpret and reflect upon the feedback.

As discussed in the beginnings of this chapter, both types of feedback can be found in the related work, yet the reasoning behind their use are not always clear. Tanveer’s user study [Tanveer et al., 2015] suggests that users preferred instructional over appraisive feedback, at least when it comes to speech modulation augmentation. However, this result is somewhat inconclusive since the study varied the information representation (textual versus symbolic) together with the scope. Schneider et al. [2015] also opted to use both instructional and

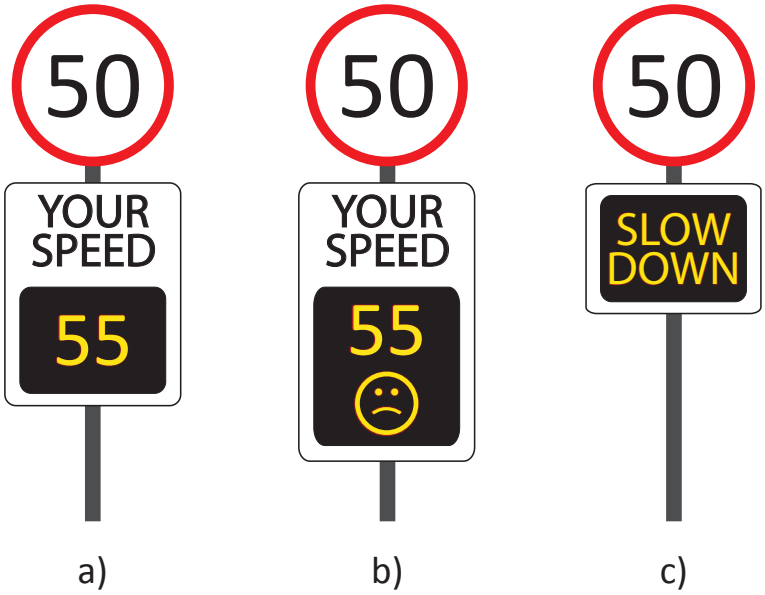

Figure 6.10: Examples of appraisive (a,b) and instructional feedback (c).

appraisive feedback in their system. However, their design choices are puzzling as they opted to use instructional feedback in response to small errors and appraisive in response to large errors. In this case, the opposite would make more sense since instructional feedback leaves less room for interpretation and thus is more suited to correct larger errors.

Both types of feedback have their advantages and disadvantages. Boyd et al. [2016] argue that appraisive feedback is better since it builds awareness and helps the user learn what they did wrong: "In this way, users can understand their own behaviours and possibly even experiment with modifications." However, the extra cognitive load induced by appraisive feedback can be disruptive to the primary task placing it at odds with the second requirement of social augmentation (feedback has a minimal impact on the primary task — see Section 4.1). On the other hand, it is expected that instructional feedback is less disruptive but its lack of transparency could confuse users and seed distrust. For example, if the augmentation instructs the user to smile more and does not provide "proof" that this is necessary, the user might lose trust in the system and ignore the feedback. Thus, from the point of view of R3 — the feedback is appropriate for facilitating a behavioural change — appraisive feedback is preferred.

6.5 Level of Detail

The final design dimension this chapter looks at is the level of detail (LoD) of the feedback. Similar to other parameters, the LoD can also be seen as a spectrum with low LoD and one side and high at the other. An example for a low LoD visual feedback is showing the text "smile more", whereas a high LoD variant would be "smile 12% more for the next 25 seconds." Similar examples can also be found for other modalities: a short vibration instructs the user to smile more (low LoD); a 12 second long vibration indicates that the user needs to smile 12% more (high LoD).

When looking at related literature, low LoD feedback is the clear favourite. Both Tanveer

Figure 6.11: The balance between R2 and R3 relative to feedback level of detail.

et al. [2015] and Boyd et al. [2016] chose simplicity over detail. A similar trend can be observed for other modalities as well [Bodnar et al., 2004; Murer et al., 2013; Suhonen et al., 2012]. Especially interesting is the use of low LoD "social" feedback by Chollet et al. [2015]. They use a virtual audience which reacts to the user with nodding and posture shifts. However, there are exceptions. Schneider et al. [2015] uses high LoD visual feedback in response to "severe" behavioural mistakes. Yet, unlike the other systems, Schneider et al.'s feedback is meant to be interruptive. The tactile feedback of van der Linden's multi-actuator setup [van der Linden et al., 2011b] can also be considered high LoD.

In terms of social augmentation, the reduced amount of information associated with a low LoD means the feedback can be perceived easier and quicker which in turn minimizes the disruption effect on the primary task. This serves the second requirement of augmenting social interactions (see Section 4.1). The downside is that low LoD feedback lacks clarity and explicitness, leaving room for interpretation. This could interfere with R3 which requires that the feedback is appropriate for facilitating the intended behavioural change. Thus, a balancing act emerges between low LoD and R2 on one side, and high LoD and R3 on the other (see Figure 6.11). Yet, it is expected that the balance is slightly skewed towards low LoD since in the realtime context of social augmentation, where feedback is delivered often and without latency, a lower LoD is likely to be sufficient.

6.6 Summary

This chapter offered an extensive overview and classification of live feedback, its properties and variations. The general aim of the chapter was to give potential designers the foundation required to make informed decision when it comes to infusing feedback mechanisms in their application and help with questions such as:

- What are the benefits of using textual feedback? (see Section 6.1.1)
- Is there a difference between the feedback "smile more" and "you don't smile enough"? (see Section 6.4)
- What options do I have for designing touch feedback? (see Section 6.1.3)
- Can you use the sense of taste for feedback? (see Section 6.1.6)

In order to achieve this, the chapter first explored related literature for examples and theories regarding feedback modalities. In this context, a total of seven variations of feedback (six unimodal and one multimodal) have been presented. The chapter then introduced four additional feedback characteristics which classify a feedback approach based on the content of the feedback messages and on how these are delivered to the user. For each modality and design characteristic, advantages and disadvantages for use in social augmentation systems have been discussed using theories from the social sciences, concrete examples of systems from the HCI domain as well as various empirical studies — three of which I have contributed

to (Sections 6.1.1, 6.1.3 and 6.2.2).

Moreover, every feedback characteristic has been positioned relative to the requirements of social augmentation postulated in Section 4.1. Figure 6.12 provides an overview of this pseudo-classification which takes the shape of a confusion matrix. Here, it is important to note that the figure has been populated using data and information from a variety of sources, including psychological theories, empirical and anecdotal findings as well as personal experience from my part. Moreover, because they cover such a wide area of research, in only the rarest of cases have the characteristics been empirically compared to one another. Thus, the figure mostly serves the purpose of summarizing the discussions of this chapter, and should not be used to compare between the individual feedback dimensions or to extract any general conclusions.

Throughout this thesis, various systems which make use of live feedback have been presented and discussed. Table 6.3 provides an overview of these systems and classifies them based on the feedback characteristics they employ. The table shows that visual feedback is very common in related literature. Furthermore, it is noticeable that some feedback categories are drastically under-represented. For example, although multiple proof-of-concepts for gustatory stimulation have been discussed, the survey revealed only one application which uses gustatory feedback in a meaningful way [Murer et al., 2013]. The table also reveals that some combinations of feedback characteristics are more difficult to achieve, if not impossible (e.g. high level of detail thermal, olfactory or gustatory feedback). Moreover, the fear of damaging the skin using thermal feedback has caused developers to be extremely careful when it comes to thermal stimulation. Thus, the surveyed systems deliver only low prominence thermal feedback.

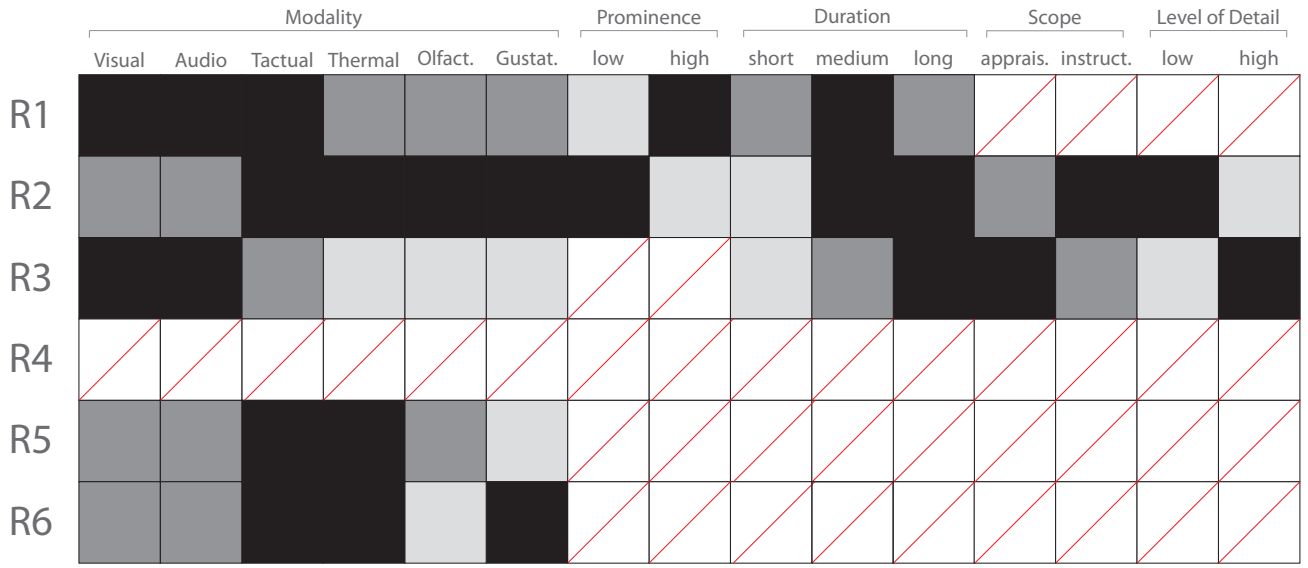

Figure 6.12: Matrix showing the relationship between the individual feedback characteristics and the requirements of the social augmentation concept. The darker a cell is, the more does the feedback characteristic satisfy the specific requirement. Four values are possible: high (black), medium (dark grey), low (light grey) and no correlation at all (barred white cell).

		Modality						
		Visual	Audio	Tactile	Thermal	Olfactory	Gustatory	Multimodal
Prominence	low	Boyd et al. 2016; Damian et al. 2015b; Müller et al. 2013; Tanveer et al. 2015			Gooch and Watts 2010; Suhonen et al. 2012; Wilson et al. 2015	Kaye 2001		
	high	Schneider et al. 2015	Damian et al. 2016a; Dietz et al. 2016	Damian and André 2016; Ofek et al. 2013; van der Linden et al. 2011b				Schneider et al. 2015
Duration	short	Boyd et al. 2016; Tanveer et al. 2015	Damian et al. 2016a; Ofek et al. 2013	van der Linden et al. 2011b	Lee and Lim 2010	Bodnar et al. 2004	Moser and Tscheligi 2013	Lee et al. 2009
	long	Damian et al. 2015b; Damian and André 2016	Banf and Blanz 2013; Dietz et al. 2016		Suhonen et al. 2012	Kaye 2001		Schneider et al. 2015
Scope	appraisive	Barmaki and Hughes 2015; Boyd et al. 2016; Damian et al. 2015b; Tanveer et al. 2015	Banf and Blanz 2013; Damian et al. 2016a; Dietz et al. 2016	Damian and André 2016		Kaye 2001	Moser and Tscheligi 2013	Schneider et al. 2015
	instruct.	Tanveer et al. 2015		van der Linden et al. 2011b		Bodnar et al. 2004		Schneider et al. 2015
LoD	low	Boyd et al. 2016; Damian et al. 2015b; Tanveer et al. 2015	Banf and Blanz 2013; Damian et al. 2016a; Dietz et al. 2016	Asif and Boll 2010; Damian and André 2016; Pielot et al. 2010	Suhonen et al. 2012; Gooch and Watts 2010; Wilson et al. 2015	Bodnar et al. 2004; Kaye 2001	Moser and Tscheligi 2013	Lee et al. 2009
	high	Barmaki and Hughes 2015; Dermody and Sutherland 2015		van der Linden et al. 2011b				Schneider et al. 2015

Table 6.3: Overview of live feedback implementations. Due to an overall lack of social augmentation systems, general HCI systems which support live feedback are also included.

III Implementation

7. The SSJ Framework

This chapter introduces the SSJ[1] software framework for building and running behavioural feedback loops, with the overall goal of augmenting human-human interactions. The framework is capable of both realtime human behaviour analysis and live multimodal feedback delivery.

SSJ has been implemented with social augmentation in mind, closely adhering to the requirements postulated in Section 4.1. In order to enable the generation of feedback that facilitates a behavioural change (R3), and which improves the user's position in the social interaction (R4), SSJ supports realtime social signal processing and advanced behaviour classification techniques. On the other hand, the framework's feedback manager offers a high degree of flexibility in designing complex feedback strategies. It supports both the delivery of subtle and undisruptive feedback, which has minimal impact on the primary task (R2), as well as highly prominent feedback capable of drawing attention from the primary task in order to facilitate a behavioural response (R1). Moreover, thanks to automatic adaptation techniques, the feedback strategy can be designed to adjust to the user over time. However, it is the fifth requirement which had the largest effect on the implementation of the framework. More specifically, in order to minimize the disruption of the social interaction (R5), the framework has been designed and built to support the execution of fully mobile and out-of-sight behavioural feedback loops using mobile devices. This enables the augmentation of in-situ social interaction with the help of just a smartphone and some untethered periphery devices (e.g. a sensor armband or a pair of headphones).

For behaviour analysis, SSJ supports a broad range of sensing hardware including smartphone internal and external sensors (see Appendix B). Moreover, various filtering and

[1]The origin of the name can be traced back to the incipient phase of the project when the intention was to build a Java equivalent of the Social Signal Processing (SSI) framework [Wagner et al., 2013] (SSI + Java = SSJ). Since then however, the project evolved beyond its initial scope but the name remained unchanged. Now, the name symbolizes both the roots of the project as well as its deep conceptual link and well defined programming interface with SSI.

feature extraction algorithms are already implemented. These, together with state-of-the-art machine learning techniques, enable the realtime classification of social signals "in the wild." On the feedback side, SSJ is able to interface with four different types of output devices spanning three modalities: visual, auditory and tactile (see Appendix C). A complete list of all implemented components of SSJ is provided in Appendix D.

SSJ is an extension of the Social Signal Interpretation (SSI) framework for Windows [Wagner et al., 2013]. It inherits the design elements which make SSI a popular choice for analysing human behaviour, and expands upon them by adding support for fully mobile social signal processing and feedback generation. To achieve this, three new design principles are introduced, which complement the initial guideline postulated by Wagner (see Section 7.1):

- *Efficient Resource Management*. Applications created with SSJ are efficient in the handling of memory, processing and energy resources.
- *Fault Tolerance*. The building and execution of social signal processing pipelines and behavioural feedback loops is robust towards user errors or connectivity problems.
- *Accessibility*. The tools and instruments required for creating and running social signal processing pipelines and behavioural feedback loops are accessible to, and usable by persons without a technical background (e.g. users of the social augmentation)

In the remainder of this section, a quick description of the original SSI framework is presented. This should give the reader a good understanding of the general concepts which lie at the centre of both SSI and SSJ. Following this, an overview of SSJ's architecture is provided and the critical design and implementation elements which make SSJ an efficient platform for executing mobile behavioural feedback loops are highlighted. This efficiency is then put to the test in a performance evaluation. Afterwards, Section 7.5 describes the feedback generation capabilities of SSJ. Furthermore, a graphical user interface which enables SSJ to be used by less technical persons, such as users of the social augmentation, is introduced. Finally, to illustrate how SSJ can be used for creating behavioural feedback loops, a tutorial is provided which guides the reader through the process of setting up a system for classifying and providing feedback in response to stress.

7.1 Origins

SSJ is based on the Social Signal Interpretation (SSI) framework [Wagner et al., 2013], from which it also derives its name. SSI is a project of the Human Centered Multimedia Lab at the Augsburg University and has been originally brought to life by my colleague Johannes Wagner. While SSI has been primarily designed to run on stationary personal computers running Windows, it can also be deployed on more portable devices such as laptops and tablet computers. Thanks to a recent port to UNIX, it can also run on smaller form factor devices but with limited functionality [Flutura et al., 2016]. One of the main goals of the framework is to facilitate the development of online behaviour analysis systems, e.g. social skills training environments. To achieve this, SSI follows four core design principles:

- *Generic Data Handling*. To support numerous heterogeneous sensors, all delivering different data types at different sample rates, SSI relies on a generic data structure for storing and manipulating data streams. This way, video data consisting of large pixel arrays updated 30 times per second, and audio data consisting of one or two values but updated several thousand times per second can be buffered, processed and classified using the same logic.

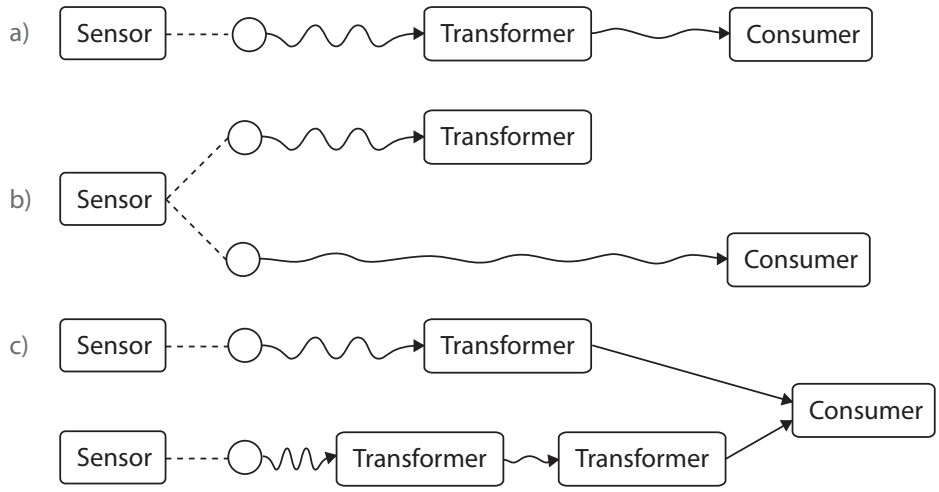

Figure 7.1: Illustration of a linear (a), a forking (b) and a fusing SSI pipeline (c).

- *Synchronization*. It is often the case that a sensor is not able to always precisely maintain its own specified sample rate and occasionally provide either too much or too little data, causing signal drift. More specifically, they occasionally provide either too much or too little data. To tackle this issues, SSI periodically performs cross signal synchronization to make sure every sensor provides the expected amount of data. Thus, the data from two different signals can be safely matched to one another. For example, if a multimodal processing pipeline requires both video and audio data to infer the emotional state of the user, SSI will maintain the data streams coming from these two sensors synchronized. This allows an event in the audio signal to be easily matched to an event in the video signal.
- *Modular Design*. Due to the rapid advancements of both sensing devices and software processing algorithms, expandability and maintenance of a social signal processing framework is crucial. SSI accomplishes this with a highly modular design which allows effortless plug-in and plug-out of individual components without affecting the functionality of the other components or the stability of the framework. The modular design also permits the splitting of long processing chains into small steps which can be reused as parts of other processing chains, thus drastically diminishing code and functionality duplication.
- *General Methodology*. In order to facilitate the expandability of the framework and to foster its growth through community involvement, a clear methodology for implementing new components or building applications is needed. SSI does this by pairing its modular design with a clearly defined API and testing suite. From an end-user point of view, the modular design allows easy "Lego-style" application building while maintaining a large degree of flexibility.

The rest of this section will explore how SSI follows these core principles by discussing its infrastructure and data flow concepts. These concepts are important since they also carry over to the SSJ framework.

7.1.1 Infrastructure

In most cases, the raw data extracted from sensors is not useful in its original state. One reason is that the data may be noisy and in need of filtering, or too broad and in need of refining. For example, the individual pixel colour values in an image say little about the facial expression of a person. First the face needs to be cut out from the image, then facial features need to be computed using image processing algorithms. Finally, these features need to be classified into facial expressions using pre-trained models. This example illustrates both the need for processing raw signals, and how such processing tasks can be split into smaller processing steps. These steps are represented in SSI as components. A chain formed out of at least two components is referred to as a pipeline.

Components are of three types: sensors, transformers and consumers. A sensor is responsible for pushing data into the pipeline. Most commonly this data is directly extracted from a physical sensor device. A single sensor can be the source of multiple signals and thus can be associated with multiple sensor channels. For example, a camera provides both audio and video data. Whereas sensors have multiple outputs, consumers have one or more inputs. This allows them to receive data from multiple components and process it simultaneously. For instance, a classifier can use both audio and video data to perform multimodal classification. Transformers have both inputs and outputs. Thus, they receive data, transform it, and then push the transformed data back into the pipeline. A typical example of a transformer is filters. They process the data stream by filtering out unwanted artefacts (e.g. noise). Another example is feature extractors, which refine the data to extract meaningful characteristics (e.g. pitch extracted from a raw audio signal).

A pipeline is a chain of components where the outputs of some components are matched to the inputs of other components. Thus, pipelines represent directed acyclic graphs where data flows along the edges and is processed in the nodes (Figure 7.1-a). Pipelines can also fork and fuse. A fork occurs when a sensor's or transformer's outputs are directed to multiple components (see Figure 7.1-b). Two pipeline branches fuse when the inputs of a transformer or a consumer come from different components (Figure 7.1-c). Moreover, two pipeline branches can also run completely in parallel, i.e. without forking or fusing. To maintain the various branches of a pipeline in sync and avoid drift, SSI regularly synchronizes each sensor channel against a central clock.

7.1.2 Data Flow

In a pipeline, data flows from component to component in the form of *streams*. A stream represents a signal window and is composed out of a finite amount of data samples. A sample is a data structure containing one or more values of a specific type. The amount of values a sample contains is called sample dimension (*dim*). For instance, a stereo audio signal contains samples of dimension 2 (one for each channel). A stream is defined by the timestamp of the first sample it is holding (*time*), the type of samples it is holding, the amount of samples per second (i.e. sample rate or *sr*) and the number of samples it is holding (*num*). This structure is illustrated in Figure 7.2.

The data streams are passed from one component to another with the help of a buffer placed between the two components. This allows the components to work independently and asynchronously of one another. Thus, components within a pipeline can work at different update rates and on different data window lengths. For example, a camera sensor can push images into the pipeline one at a time whereas an image classifier, which follows it in the

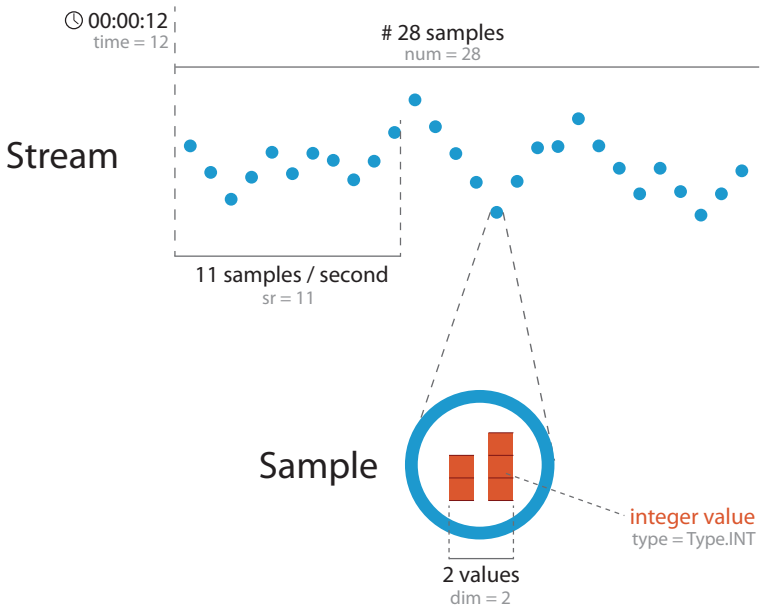

Figure 7.2: Illustration of the stream and sample concepts and their properties.

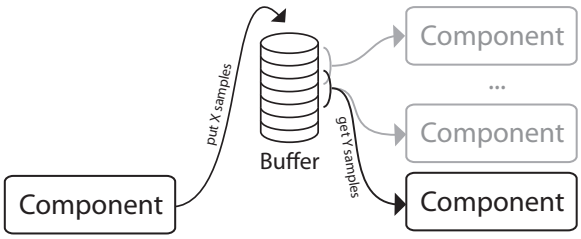

Figure 7.3: Buffer-backed data flow through a pipeline.

pipeline, can work on several seconds worth of data at once. Moreover, while only one component is able to push data into a buffer, multiple components are able to simultaneously read from it, allowing forks in the pipeline. The data flow concept is illustrated in Figure 7.3.

Besides streams, data can also be passed between components as *events*. Whereas streams are windows of a larger continuous signal, events are discrete. They lack a sample rate, a fixed duration as well as any regularity. The role of events is to represent discrete and irregular occurrences which are difficult to shape into a continuous signal (e.g. utterances or gestures). Events are passed from component to component using the observer pattern as illustrated in Figure 7.4. More precisely, every component can register as a listener for a particular event type. As soon as another component sends out a matching event (Figure 7.4: steps 1 and 3), all listening components will get notified (Figure 7.4: steps 2 and 4). This logic is handled by a central event board, which acts similarly to the buffers used for stream-based communication.

7.1.3 Summary

The concepts and design principles introduced in this section also lie at the heart of SSJ. This makes SSJ an easy expandable software solution for performing multimodal signal

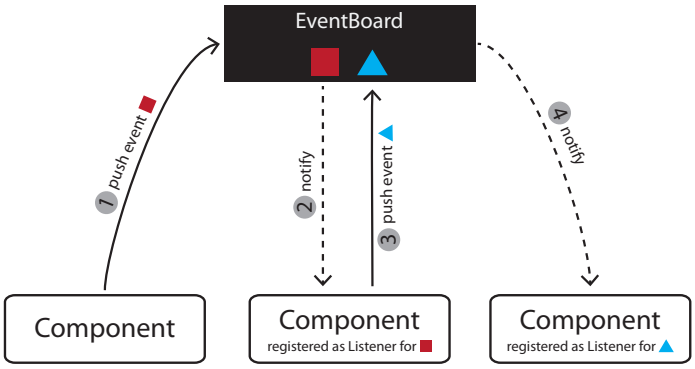

Figure 7.4: Observer pattern for event-based communication between components.

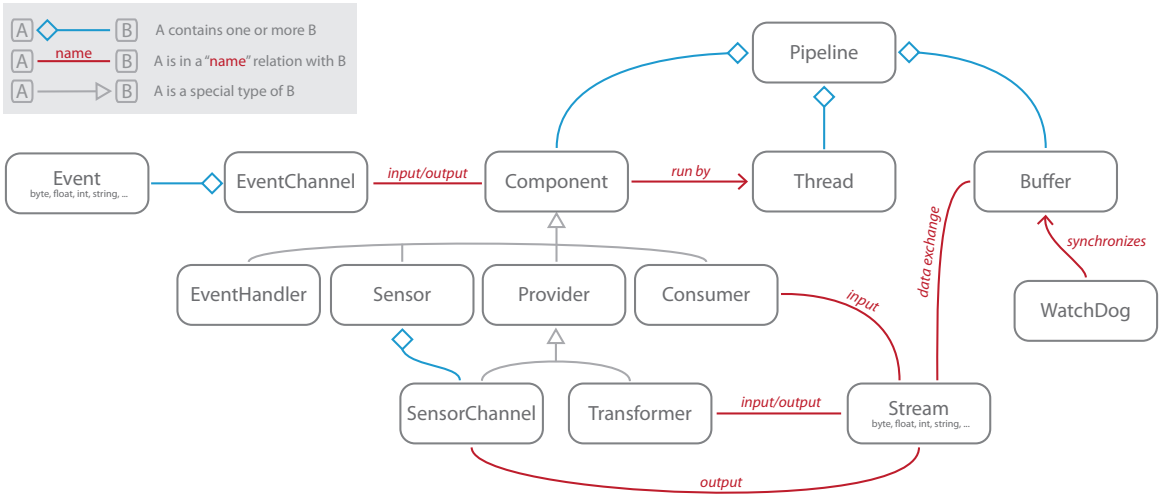

Figure 7.5: A class diagram of SSJ's core.

processing in a mobile environment. The following paragraphs will look into more detail on how this is achieved. For more information on the implementation of the original SSI framework, the interested reader is encouraged to read Wagner's Dissertation [Wagner, 2016].

7.2 Architecture

On account of their strong conceptual link, the architectures of the original SSI and the mobile SSJ signal processing solution bear various similarities. However, due to the shift to the Android ecosystem and a different programming language, SSJ's architecture differs in several key areas. Thus, a quick look at the overall architecture of SSJ is needed to lay the groundwork for the discussions in the following sections. Figure 7.5 provides an overview of the architecture of SSJ's core.

The main class of SSJ is the *pipeline* class. It is the representation of a social signal processing pipeline and is responsible for managing the components, threads and buffers of the pipeline. It also houses the public programming interface for configuring and executing a pipeline. Thus, it offers several functions for adding various types of components to the pipeline as well as starting and stopping it.

A *component* represents a processing step in the pipeline. Its execution is handled by a dedicated *thread*. There are five types of components: *sensor*, *sensor channel*, *transformer*, *consumer* and *event handler*. The sensor is responsible for managing the connection to a physical or virtual sensor device. A single sensor can provide multiple signals using different sensor channels (e.g. the Microsoft Band 2 provides both heart rate and skin temperature measurements). Transformers and consumers are analogue to those in SSI and are responsible for data processing. Sensor channels and transformers are also referred to as *providers*.

For each provider added to the pipeline, a *buffer* is created to manage the data it produces. Thus, there is a buffer between every two connected components. Every buffer acts as a sink for one component and as a source for one or more components (see Figure 7.3). Signal windows are extracted from the pipeline's buffers and provided to the components in form of input *streams*. Once a component finishes its processing, it places its data in an output stream that is then copied into the pipeline buffers, allowing other components to access it. Sensor channels and consumers have only input or output streams, transformers have both.

At runtime, SSJ synchronizes sensor signal with the help of *watch dogs*. These continuously monitor the buffers of all registered sensor channels and make sure they are being written to with a correct sample rate.

Discrete communication between components is carried out with the help of *events*. Generally, every component in SSJ can work with events. However, the *event handler* is a special type of component which lacks any input or output streams and thus can only communicate using events. Unlike SSI's central event board, the event data flow in SSJ is handled by individual *event channels*. A component can register itself as a listener to an event channel and be automatically notified when a new event is available. Events are stored in the event channel in a first-in-first-out manner.

7.3 Going Mobile

At the core of SSJ's mobility lies its choice in target platform: Android. The mobile operating system was initially developed by Android Inc., a company which was acquired by Google in 2005. The operating system was first unveiled in 2007 with the goal of providing an open source alternative to the closed and restrictive products of other companies. This openness has led to a fast adoption rate which quickly transformed Android into the most widely used operating system for smartphones. At the time of this writing, Android is powering over 2.1 billion units worldwide, which accounts for roughly 80% of all smartphones in the world[2]. However, Android is also used to power a variety of other mobile devices including tablets, smart glasses and smart watches. The versatility and ubiquitousness of Android makes it an ideal platform for augmenting social interactions.

SSJ has been developed from the ground up for the Android ecosystem. Thus, SSJ is able to run on virtually all Android devices (API 16 or newer) including smartphones, tablets, smart glasses (e.g. Google Glass, Lumus DK-40, Epson Moverio) and smart watches (e.g. Moto 360, Samsung Gear), allowing an unprecedented flexibility and mobility for performing social signal processing.

This section will discuss the key elements which enable SSJ to perform complex social signal processing tasks on such a large variety of devices. First, the particularities of running

[2] https://www.statista.com/statistics/385001/smartphone-worldwide-installed-base-operating-systems

SSJ on the Android platform are presented. Afterwards, a detailed discussion on the resource management, energy efficiency and fault recovery mechanisms is provided.

7.3.1 Adapting to the Android Platform

In order to run on a large amount of devices, each with different hardware configurations, SSJ has been implemented entirely in Java. This allows SSJ to be independent of the hardware specifications of the individual devices. However, Java's higher level programming perspective meant that translating some of SSI's more intricate mechanisms was not always easy. The generic data handling of SSI, which allows the entire framework to manipulate data structures independent of their type, is one example. For this, it relies on *C++*'s versatile pointer mechanic which allows it to freely cast between data types. For example, a *byte* array can be treated as a *float* array by simply casting the pointer from *byte** to *float**. However, Java's lack of access to object pointers means that one cannot cast primitive arrays. A different solution was needed. SSJ uses an object oriented approach consisting of a generic *Stream* class which is extended by a series of data type specific subclasses. However, the internal buffers which manage data flow between components are type invariant and implemented as byte arrays. To facilitate copying between buffers and stream objects, a versatile set of efficient array copy functions have been implemented. They enable copying data from *byte* arrays to any other type and back, offering a similar functionality to *C++*'s *memcpy()*. Informal tests have shown that these type-invariant array copy functions are only marginally slower than Java's own type-variant *arraycopy* mechanic. Yet, they allow SSJ to internally handle the various signals in a generic, data type invariant fashion while the components can work normally with typed data structures.

The switch to Java also brought improvements in terms of code structure and readability. Complex and convoluted *mutex* mechanics for regulating multi-threaded workflow have been replaced with simple synchronization blocks. To further improve readability, maintainability and expandability, a stronger object oriented perspective has been employed. Streams are now objects which are aware of what data they contain. This means that at any point in a pipeline, it is possible to know not only the type of data a stream contains, but also its origin and what its content represents. For example, unlike SSI, a transformer can actually verify if the input stream is video coming from a camera or acceleration coming from an inertial measurement unit. The opposite is also true, every component can trace back the streams it created. This allows SSJ to drop pipeline connection objects (SSI's *transformables* and *pins*). Now, to connect two components in a pipeline, the component instances themselves can be used, simplifying pipeline code structure.

7.3.2 Managing Performance

A software program consumes two types of resources: processor cycles and memory. As discussed in Section 5.1.3, mobile devices are generally more limited in terms of these resources. This combined with Java's high level perspective and tendency to sacrifice efficiency in favour of code readability, posed a sizeable challenge for SSJ.

One big difference between C++ and Java in terms of resource management is Java's garbage collector. The garbage collector is an automated process which periodically verifies the state of the memory. If it finds memory cells which have been written to, but are currently no longer in use, it will free them and make them usable again. The advantage of the garbage collector is that it saves the programmer from performing the unattractive task of memory

management. However, the disadvantage is that in order for it to analyse memory usage, it needs to momentarily pause the program, wasting processor cycles. Thus, whenever the garbage collector starts, SSJ's processing threads are briefly paused. While this may not be particularly damaging to a standard Android application, it can cause significant data loss for a signal processing task. For example, if the garbage collector pauses a camera sensor thread for 30 milliseconds every second, the pipeline would miss one image every second (assuming the video stream has a sample rate of 30). In certain scenarios, this could cause the system to overlook important behaviours (e.g. saccades or blinking). Moreover, the garbage collector-induced pauses diminish the time components have for data processing, potentially impacting the ability of the pipeline to run in realtime. Thus, an efficient memory management which aims at minimizing the frequency and severity of garbage collection interventions, also leads to a more effective use of processor cycles. To achieve this, several programming guidelines have been defined:

- *Do not allocate memory after the initialization phase of a pipeline*. A component has to allocate sufficient memory for any variables it will use at runtime either in the constructor or at one of SSJ's pre-runtime initialization points.
- *Avoid using functions or operations which implicitly allocate memory*. A typical example here is string concatenation which creates a new string object for every concatenation operation. This can be avoided by using the *StringBuilder* class.
- *Use memory structures which are appropriate for their task*. For example, *LinkedLists* are more efficient for iterating through them, whereas *ArrayLists* are more efficient for directly accessing the individual elements.
- *Do not use the long and double data types unless absolutely necessary*. Java consistently allocates float and integers four bytes of memory whereas doubles and longs receive eight bytes. Since it is common is SSJ to work with large data arrays, it is important that these data arrays are not larger than they need to be.
- *Avoid copying or duplicating data*. When moving data from component to component, SSJ automatically creates local copies of the signal windows. Thus, the individual components can freely work on the provided data structures and should not create additional buffers.

These guidelines allow SSJ to perform various online social signal processing tasks normally reserved to powerful personal computers: data classification, complex feature extraction, video processing operations and others.

7.3.3 Energy Efficiency

Unlike personal computers, mobile devices lack a continuous power source and rely on integrated, finite batteries. Thus, energy efficiency is a critical aspect for mobile social signal processing (see Section 5.1.3). To increase energy efficiency, SSJ makes use of Android's built-in energy saving techniques. More specifically, SSJ dynamically controls the device processor's power state in an effort to minimize the amount of time it is kept in a waking state. To achieve this, SSJ acquires a CPU wake lock just before components commence data processing and releases it immediately afterwards. This way, the processing power of the device is optimally allocated. Moreover, the Android operating system can put the component's thread in a sleeping state while it waits for new data to arrive. Thus, an application, which favours battery life over latency, can achieve this by reducing the update of the components. The use of CPU wake locks also allows the display of the device to be turned

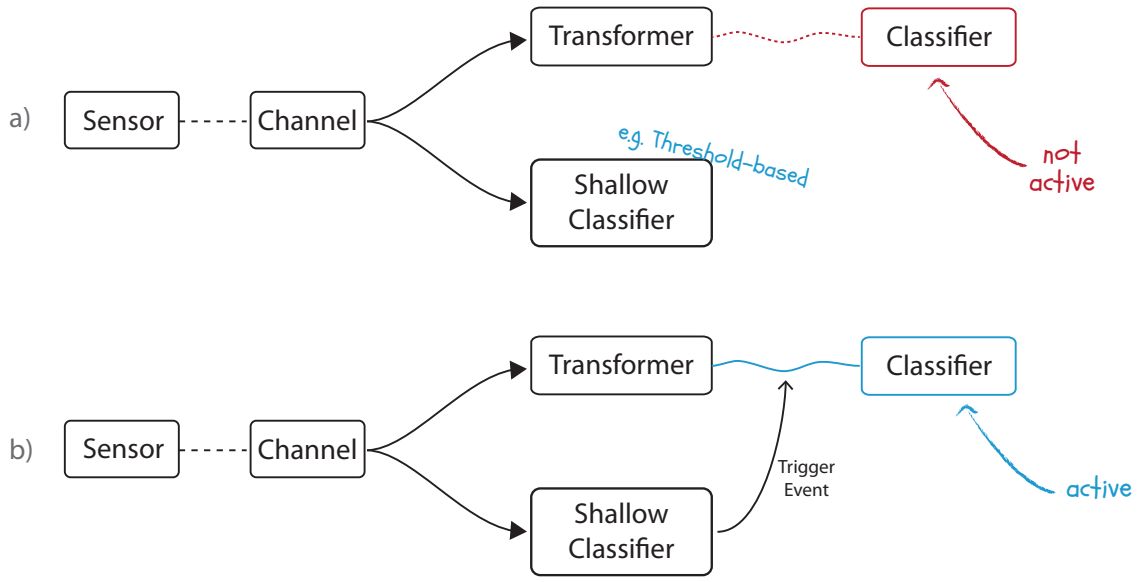

Figure 7.6: Event trigger mechanism for dynamically turning a classification step on (b) and off (a).

off during processing to maximize battery life but without impacting pipeline functionality and accuracy.

Similarly to other mobile signal processing systems [Lu et al., 2010; Rachuri et al., 2010; Wang et al., 2009], SSJ is also able to use triggers to turn on and off processing steps. This is done using events to control when a consumer should process its data. An event is a discrete message which can be sent from any component, and is commonly used to mark the occurrence of a particular behaviour such as speech or gesture. It contains both a description of the occurrence as well as timing information which specifies when, and for how long the occurrence was observed. As soon as a consumer receives an event, it wakes up and requests a data window corresponding to the timing of the event. As soon as the processing is done, the component returns to a sleeping state until a new event is received. Thus, the consumer is only active when it receives an event. This mechanic can be used to create energy efficient pipelines. For instance, a speech classifier can be configured to only become active when the user speaks. In this case, the speech detection would be performed by a more energy efficient shallow classifier (e.g. a simple threshold-based classifier). Figure 7.6 illustrates this mechanism.

Another large source of energy consumption is communication with external devices, especially sensors. To minimize the impact on battery life, SSJ makes use of the Bluetooth Low Energy API for communicating with most sensors. Bluetooth Low Energy, also called Bluetooth 4.0 LE, is an evolution of the standard Bluetooth wireless technology and offers considerable improvements in energy efficiency.

7.3.4 Fault Tolerance

To make building and running social signal processing pipelines easier and more transparent, SSJ offers an improved error handling subsystem that allows pipelines to often times recover from minor errors (e.g. caused by faulty pipeline configuration, connectivity issues) without

system crashes. This is achieved with the help of Java's Exception mechanic. Any unhandled errors or exceptions are caught by SSJ and can be forwarded to the application in charge of managing the pipeline. The application is then able react to the error, e.g. safely shut down the pipeline to prevent data loss and alert the user.

An additional fault tolerance mechanic has been implemented for sensor-type components. The connection of a sensor is periodically tested by SSJ. In case a connection is interrupted, SSJ will automatically attempt to reconnect. For example, if the smartphone moves out of range of a Bluetooth-connected sensor, SSJ will restart the connection process. This would cause the pipeline to reconnect to the sensor once it is back in range. While the connection is interrupted, the internal SSJ buffer monitoring mechanisms would take over and insert dummy values into the pipeline (i.e. zeroes) to not disrupt the functionality of the other components. This way, in a pipeline with multiple sensors, connection issues with one sensor will not lead to a system wide failure. For example, in a multimodal classification scenario, a robust classifier could still make inferences about the user's behaviour even if one sensor disconnects.

7.3.5 Performance Measurements

In order to give the reader a better impression of how well SSJ can tackle typical social signal processing tasks, a demo pipeline has been implemented and evaluated for performance and energy consumption. This test should (a) give the reader an impression of the computational power and battery life of mobile devices when performing social signal processing tasks, and (b) demonstrate the potential of SSJ to perform "in the wild" behaviour analysis.

Pipeline

For the testing, an SSJ classification pipeline has been configured. First, a *FileReader* sensor is used to read a pre-recorded audio file from the SD card. The sound file in question is 26.72 seconds long and has been recorded at 16000 samples per second.

```
// Setup
Pipeline pipe = Pipeline.getInstance();

// Sensor
FileReader file = new FileReader();
file.options.fileName.set("audio.stream");
FileReaderChannel audio = new FileReaderChannel();
pipe.addSensor(file, channel);
```

The audio data stream is then processed using two transformers. The *FFTfeat* transformer computes the fast Fourier transform (FFT) coefficients for every 512 audio samples. For every coefficient, 11 common functionals are computed including mean, energy, minimum, maximum and others. This results in a feature vector of size 2827. The functionals are computed once every 0.5, 1.0 or 5.0 seconds (sample rate 2.0, 1.0, or 0.2 Hz). This parameter was varied during the evaluation to test different CPU loads.

```
// Transformers
FFTfeat fft = new FFTfeat();
pipe.addTransformer(fft, audio, 512.0 / audio.getSampleRate(), 0);

Functionals func = new Functionals();
pipe.addTransformer(func, fft, X, 0); //X = 0.5, 1.0 or 5.0
```

The data stream is then fed into a *ClassifierT* transformer. It uses an SVM model which has been previously trained in SSI with 780 samples. The aim of the model is to classify chunks of a short speech as either belonging to the introduction or to the conclusion. The classifier uses the same update rate as the *Functionals* transformer.

```
// SVM-based classification
ClassifierT classifier = new ClassifierT();
classifier.options.trainerFile.set("svm.trainer");
pipe.addTransformer(classifier, func, X, 0); //X = 0.5, 1.0 or 5.0
```

Finally, the output of the classifier is printed on the console using the *Logger* consumer.

```
// Consumer for console output
Logger log = new Logger();
pipe.addConsumer(log, classifier, X, 0); //X = 0.5, 1.0 or 5.0
```

Resources

The pipeline has been tested on two smartphones: Samsung Galaxy S4 (GT-I9505) and Huawei Nexus 6P. The Samsung S4 is powered by a quad-core Snapdragon 600 CPU, running at 1.89 Ghz, as well as a 2600 mAh battery, which offers 9.88 Wh of energy. The operating system is Android version 5.0.1. Due to its relative age (it was released in 2013) and moderate specifications, the Samsung S4 is a good representation of an entry-level smartphone. In contrast, the 2016 Huawei Nexus 6P is a high-end smartphone powered by an octa-core Qualcomm Snapdragon 810, running at 1.96 GHz, and a 3450 mAh battery with an estimated 13.29 Wh. On the software side it runs an unmodified version of Android 7.0. For the test, both phones have been reset to factory settings.

To measure energy consumption rate and CPU load, the Qualcomm Trepn Profiler[3] has been used[4]. The software's advanced sensing features and battery power estimation algorithm make it ideally suited for this test [Hoque et al., 2016]. Trepn has been chosen over the more popular, but older, PowerTutor[5] profiler due to better compatibility with modern smartphones.

Method

Two tests have been performed. First, three versions of the pipeline have been run on both devices for 120 seconds. The versions differed in processing and classification sample rate. More specifically, the sample rate of the *Functionals* and *ClassifierT* transformers was varied between 0.2 Hz, 1.0 Hz and 2.0 Hz. This test was meant to show how much power the SSJ pipeline draws at different configurations. To test the total uptime, the pipelines were started on fully charged devices and left to run until the battery reached 5%.

The second test aimed to find the maximum classification sample rate for each device. To this end, the same pipeline was executed multiple times with ever greater sample rates. This process was stopped as soon as pipeline errors started to occur. More specifically, when the sample rate is too high, the feature extraction and classification threads would no longer finish processing a sample before the next one arrives, creating a bottleneck. If the bottleneck persists, the resulting processing latency can become too great for SSJ's buffers, causing samples to be dropped.

[3] `https://developer.qualcomm.com/software/trepn-power-profiler`

[4] The Samsung Galaxy S4 is not officially supported by Trepn, however, preliminary tests yielded credible values.

[5] `http://ziyang.eecs.umich.edu/projects/powertutor`

Device	SR	Power (W)	CPU (%) 1	2	3	4	Battery Life (h:m:s)
Samsung S4	2.0	1.131	67.00	66.87	64.68	61.40	10:06:24
	1.0	0.796	62.39	64.09	62.90	59.83	11:54:03
	0.2	0.745	59.82	61.89	61.52	53.39	13:55:23
Nexus 6P	2.0	0.251	44.89	36.40	30.63	9.54	18:53:07
	1.0	0.223	46.02	37.09	30.90	8.39	24:13:49
	0.2	0.227	47.65	38.51	31.87	8.52	30:48:34

Table 7.1: Mean energy consumption rate, CPU load (for each core) and battery life.

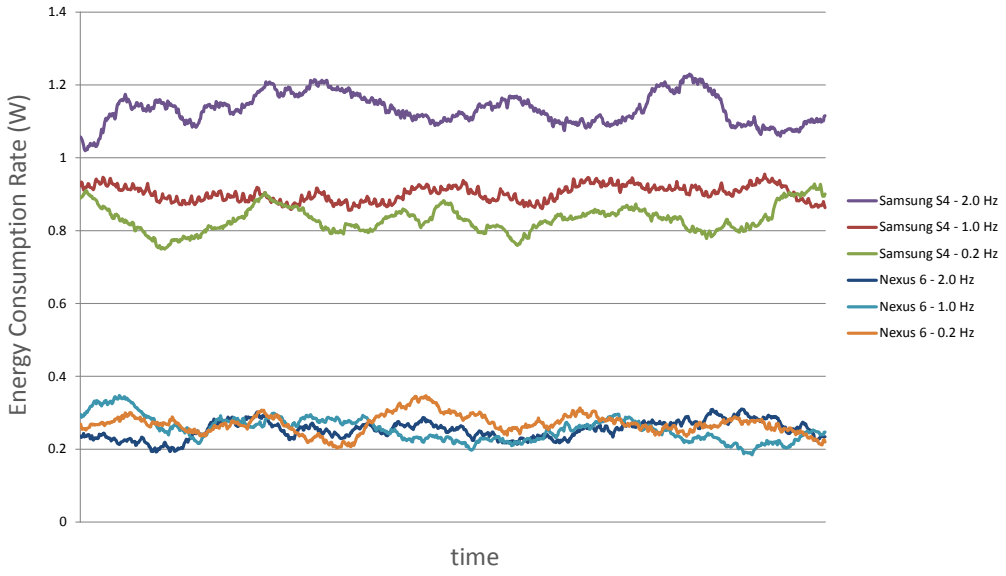

Figure 7.7: Energy consumption rate over time between devices and pipeline configurations.

Results and Discussion

The results of the first test are summarized in Table 7.1. Figures 7.7 and 7.8 also provide an illustration on how the energy consumption rate and CPU load varied over time. The energy consumption values are reported relative to the consumption measured before the start of the pipeline. Thus, the indicated values can be read as how much power the SSJ pipeline requires, independent from the consumption of the other processes running on the smartphone. The CPU load presented in Table 7.1 is reported by cores. However, it must be noted that although the Nexus 6P's CPU has eight physical cores, they are split in two blocks of four (one for performance, one for energy efficiency) and only one block can be active at any given time. The table reports the loads of the four cores from the performance block which was active during the test.

The generally small consumption rates demonstrate the energy efficient nature of SSJ, with uptimes of over 24 hours (for the Nexus 6P). The reported uptimes also show that the energy consumption values provided by Trepn are not perfectly accurate. More specifically, the theoretical uptimes achieved by dividing the battery capacity of each phone by the reported energy consumption rate differ from the ones achieved through testing[6]. In particular, the

[6]Since the values reported in Table 7.1 represent only the consumption rate of the SSJ pipeline, the standby

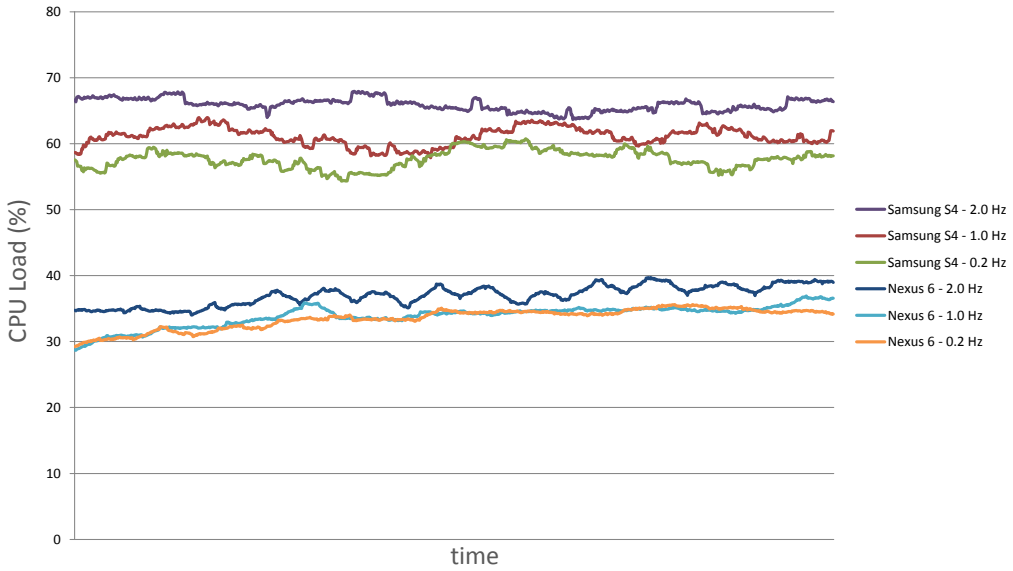

Figure 7.8: Combined CPU load over time between devices and pipeline configurations.

differences between the individual configurations are much more pronounced for the battery lives yielded by the actual test.

Looking at Figure 7.7, it becomes clear that the Samsung smartphone was far less efficient than the Nexus. However, this is to be expected since the greatest focus for CPU developers over the recent years was improving energy efficiency. This efficiency also translated to memory management, with the pipeline requiring an average of 55 MB of memory on the Samsung S4 and only 33 MB on the Nexus 6P. It is likely that this can be attributed to improvements in memory management in Android 7. Moreover, the performance gain of the Snapdragon 810 over the Snapdragon 600 allows the Nexus 6P to perform the same computations faster resulting in an overall smaller CPU load.

When comparing between the different configurations, we see that the Samsung S4 consumed 34.13% less energy and had a 37.72% longer battery life when the classifier was running at 0.2 Hz compared to 2.0 Hz. For the Nexus 6P, the energy consumption values reported by Trepn are less clear (all configurations had a similar consumption), suggesting an error in its estimation model. However, looking at the battery uptime we notice a trend similar to the Samsung device: The Nexus 6P had a 63.13% longer uptime in the 0.2 Hz configuration compared to the 2.0 Hz configuration. Thus, as predicted in Section 7.3, sample rate does impact the energy consumption of the pipeline.

The second test yielded a similar picture in terms of differences between devices (see Table 7.2). Whereas the Samsung S4 reached a maximum classification sample rate of 3.906 Hz, the Nexus 6P easily surpasses this value and manages 15.625 Hz without errors.

energy consumption needs to be added (roughly 0.4 W for both phones with connectivity services disabled). This leads to theoretical battery lives of 6.45/8.26/8.63 hours for the Samsung S4 and 20.42/21.34/21.21 hours for the Nexus 6P.

Device	SR	Errors
Samsung S4	3.125	no
	3.906	no
	4.464	yes
Nexus 6P	7.813	no
	15.625	no
	31.250	yes

Table 7.2: Stability at high sample rates during classification task.

7.4 Interfaces

Besides mobility, the second defining characteristic of smartphones is connectivity. They offer various communication interfaces, including WiFi, Bluetooth or NFC. SSJ taps into this connected world to facilitate distributed signal processing setups spanning multiple devices.

The primary cross-device interface is SSJ's Bluetooth input/output gateway, which allows pipelines to send and receive continuous signals and discrete events to and from other devices. Sending devices make use of the Bluetooth writer components, which employ Android's Bluetooth sockets for data transmission. If added to a pipeline, a Bluetooth writer will forward SSJ data streams over the Bluetooth connection to other devices. One writer can send multiple data streams at the same time using the same Bluetooth connection. This minimizes the amount of necessary Bluetooth connections and reduces the energy footprint of SSJ. On the receiving end, Bluetooth reader components are used to receive data and push it into the pipeline. One reader can also handle multiple data streams at once. As discussed above, SSJ is tolerant towards connection interruptions. If a Bluetooth connection is lost, both the reader and writer will actively attempt to re-establish the connection. Two types of cross-device Bluetooth connections are supported: stream and event-based. Stream-based connections are handled by the consumer *BluetoothWriter* and the sensor *BluetoothReader*, whereas event-based connections are handled by the components *BluetoothEventWriter* and *BluetoothEventReader*. As discussed before, SSJ is also able to handle Bluetooth Low Energy (LE) connections from various external devices. Most sensors have a dedicated sensor class which handles the BLE connection. However, SSJ also includes a generic *BLESensor* which continuously listens for BLE messages with a user-defined service and characteristic, and forwards them into the pipeline as samples in a data stream.

Besides Bluetooth, SSJ is also able to communicate with other devices over WiFi. The functionality of SSJ's WiFi connection is analogue to that of the Bluetooth connection, meaning that both stream and event-based connections are supported. However, the large benefit of the WiFi interface is that it allows SSJ pipelines to also communicate with SSI pipelines running on desktop computers.

When running a distributed signal processing setup, with multiple devices running individual interconnected pipelines, keeping the pipelines synchronized with each other is critical. It makes data streams comparable across devices. This is particularly crucial for mobile devices where aggressive energy management routines in the devices' operating systems may cause fluctuations in system clock accuracy. To address this, SSJ employs a two-step strategy to keep the pipelines synchronized. First, a master pipeline can be defined which, at startup, broadcasts a start signal over WiFi network to all other devices. This allows all pipelines to start at the same time. The second cross-device synchronization mechanic

incorporated in SSJ is a continuous clock sync. This allows multiple pipelines to periodically synchronize their internal clocks against the clock of the master pipeline. For this, Cristian's clock synchronization [Cristian, 1989] algorithm has been implemented. More specifically, all pipelines in a distributed setup will regularly request for the master pipeline's internal clock T. Upon receiving T, the slave pipelines will set their local clock to $T + \frac{RTT}{2}$. RTT is the round-trip-time, i.e. the time which has elapsed between sending out the message and receiving the response from the master. To increase accuracy and diminish the impact of WiFi quality fluctuations, every request is sent out multiple times and only the response with the smallest RTT is used. Using this strategy, distributed signal processing tasks can be carried out on multiple devices without worry of synchronization loss.

SSJ is also able to record a data stream on the device's memory (internal or SD card). This is accomplished using writer components such as the *FileWriter*, *AudioWriter*, *WavWriter*, *CameraWriter* or *FileEventWriter*. Data which has been recorded by SSJ is compatible with the SSI format. Thus, it can be read and processed by SSI pipelines.

Finally, realtime data exchange within the same process is also supported. Extracting data from a pipeline can be done by implementing SSJ's *EventListener* interface and registering as a listener to a particular event channel. For example, this method can be used by a graphical user interface to display the results of a classification to the user. Inserting data into a pipeline can be accomplished by simply pushing new events into an event channel to which an SSJ component is registered as a listener. For instance, information from input fields can be inserted into the pipeline to be stored on the SD card together with the sensor data in a consistent format.

7.5 Feedback Manager

To enable the design and execution of custom behavioural feedback loops for use in social augmentation scenarios, SSJ includes a feedback management component. It supports the automatic generation and delivery of multimodal feedback using various output devices (for a complete list of supported output devices, see Appendix C).

This section will first go over the general architecture of the component before delving into the details of each implemented modality. For each modality, examples and potential use cases are discussed. Following this, the section discusses the feedback manager's ability to automatically adapt the feedback to the user's reaction.

7.5.1 Architecture

The *FeedbackManager* component is implemented as an SSJ *EventHandler*. Thus, it can be added to a pipeline as a listener for behaviour analysis events. Once the feedback manager receives an event from the pipeline, it is first matched against a preconfigured feedback strategy (specified in the component's options). If the event satisfies the conditions of the strategy, feedback is triggered using an output device.

The feedback strategy is split into multiple feedback classes. Each feedback class defines what feedback to be sent in response to which behaviour. The feedback strategies are defined using XML. This allows users without extensive programming knowledge to easily design them using a text editor. The structure of the XML is analogue to the component's architecture. The feedback strategy (`<strategy/>`) can contain one or more feedback classes (`<feedback/>`). Each class defines exactly one behaviour condition (`<condition/>`) and one or more feedback

Figure 7.9: Examples of visual feedback: single icon appraisive feedback (a), double icon appraisive feedback (b) and instructional feedback (c).

actions (`<action/>`) which will be triggered whenever the specified condition is fulfilled. Each condition is linked to an SSJ event as specified by the *event* and *sender* attributes. Upon reception, the feedback manager parses the SSJ event and extracts a single value from it. If this value lies within the specified thresholds (attributes *from* and *to*), the condition is considered satisfied and the specified action is triggered. The feedback manager also contains some special condition types which use custom event parsing routines (e.g. SpeechRate, SpeechDuration, Loudness). These are defined in the XML using the *type* attribute. The action to be executed can be configured using various attributes (e.g. image source for visual feedback events, vibration intensity for tactile events). The full XML schema is shown in Appendix A. A simple feedback strategy is illustrated below, whereas more complex strategies are included in Appendices F and G.

```
<strategy>
 <feedback type="tactile">
  <condition event="energy" sender="ssj" from="0.0" to="0.8"/>
  <action duration="500" intensity="70"/>
 </feedback>
</strategy>
```

In this example, the feedback manager will trigger a vibrotactile feedback (i.e. vibration) with a duration of 500 ms and an intensity of 70, every time an *energy* event with a value between 0.0 and 0.8 is received.

7.5.2 Modalities

Three feedback modalities are currently implemented: visual, auditory and tactile. Thanks to the modular architecture, further modalities can be easily implemented once capable output devices become available. The system can be configured both unimodally (only one feedback class is defined in the configuration XML) or multimodally (two or more distinct feedback classes are configured). In a multimodal setup, SSJ supports all configuration for multimodal interaction (see Section 6.1.7).

Visual

The visual feedback class allows the system to display custom images on an Android device's screen. Up to two distinct images can be shown by a feedback action. This allows the modelling of both appraisive (one image represents the current behavioural state, the other one

is its interpretation) and instructional feedback (gives instruction on how to change behaviour – see Section 6.4). The images are defined in the XML configuration file with the help of the *res* attribute. To load the images, the feedback manager uses Android's *ImageSwitcher* interface. This allows images to fade in and out on feedback activation or deactivation. The fading duration can be defined using the *fade* attribute in the XML configuration file. Figure 7.9 provides some visual feedback examples.

Depending on the output device's screen size, multiple visual feedback classes can be shown at the same time. To achieve this, the graphical layout is dynamically generated at startup based on the loaded strategy. More precisely, a *TableLayout* is populated with as many columns as there are visual feedback classes in the configuration file. Each column has either one or two rows, depending on how many images the feedback actions contain.

The duration of the feedback can be customized with the XML attribute *duration*. From an implementation point of view, the duration represents a timer after which the *ImageSwitcher* view is cleared. Setting the duration to 0 effectively makes the feedback persistent. The prominence of the visual feedback can also be controlled using the *brightness* attribute. This will adjust the brightness of the screen while the feedback action is displayed.

Auditory

Auditory feedback is supported using any speaker or headphones connected to an Android device by wire or Bluetooth. The auditory feedback consists of an audio file (XML attribute *res*) which is played back whenever the condition is satisfied. The prominence (loudness) of the feedback action can be configured as a value between 0 and 1 using the XML attribute *intensity*. Audio playback is handled by Android's *SoundPool* interface, configured to run on the notification stream. This allows the audio resources to be executed fast and with only minor latency.

Tactile

Vibrotactile feedback is currently supported through the use of a Myo armband or a Microsoft Band 2. The Myo is usually worn on the forearm about 3 cm below the elbow, however due to its elasticity it can also be fitted on the wrist or lower leg. Figure 6.4 shows a user wearing the Myo on the forearm. Using the official Myo Android API v0.10.0 and custom GATT commands, SSJ is able to execute customizable vibrations on the armband. More specifically, both the duration of the vibration in milliseconds (XML attribute *duration*) and its prominence (intensity) – value between 0 and 255 (XML attribute *intensity*) – can be defined. Simple vibrotactile patterns can also be executed by specifying the duration and the intensity of the feedback action as an array. For example, a tactile feedback with the duration $[500, 1000, 1500]$ and the intensity $[50, 100, 150]$ would cause a progressive vibration pattern, where the vibrations gradually get longer and more intense.

The Microsoft Band 2 is worn similarly to a watch on the left or right wrist. When using the Microsoft Band 2, nine pre-defined types of vibrations (XML attribute *type*) can be executed. These range from single low intensity pulses to multiple high intensity pulses and even progressive vibration patterns.

7.5.3 Timing Management

SSJ is also able to manage the timing of feedback actions to reduce the frequency of feedback-induced interruptions of the primary task. More specifically, a lock period, which restricts the execution of other actions for a certain amount of time, can be defined for each feedback

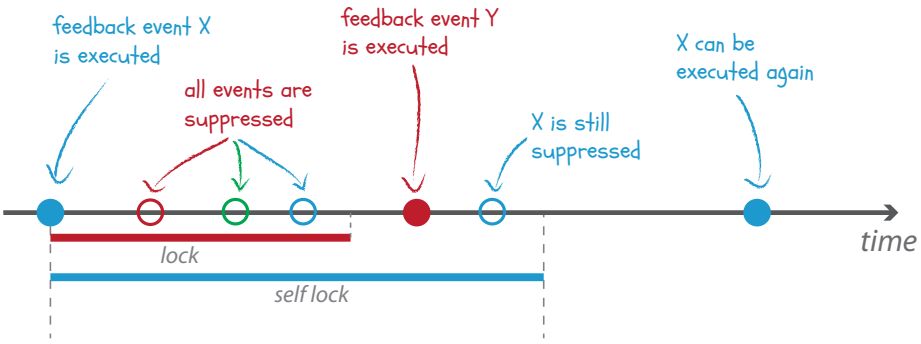

Figure 7.10: Timing management using the lock mechanic. It allows the definition of time windows during which actions of the same modality are suppressed.

action. There are two types of locks: a global lock (XML attribute *lock*) that blocks all actions of the same modality from execution, and a self-lock (XML attribute *lockSelf*) which only blocks the execution of the same action instance (see Figure 7.10).

For example, setting the global lock to 5000 ms and the self lock to 10000 ms would mean that up to five seconds after the execution of an action X, all actions of the same modality are suppressed, between five and ten seconds afterwards only the X action is suppressed, and after 10 seconds all actions can be executed.

7.5.4 Feedback Adaptation

SSJ implements the online adaptation mechanics introduced in Section 4.3.2. More specifically, a feedback validation and a feedback level system are implemented.

To enable feedback validation, feedback classes are split into two types – desirable and undesirable – according to the valance of the associated behaviour. The desirability of a feedback class is manually defined in the strategy (XML attribute *valence*). For example, a feedback which triggers whenever the user's speech rate exceeds 4 syllables per second should be defined undesirable. The overall valence of the system is considered undesirable if its last executed feedback class was undesirable. This allows SSJ to validate the impact of the feedback actions. More specifically, if a feedback action has a positive effect on the user's interaction, the valance of the system would switch from undesirable to desirable. If this does not happen, then the feedback either has no impact on the quality of the user's behaviour or it actually damages it.

In order for SSJ to be able to react to the feedback validation, a feedback strategy can be split into different levels. This is defined using the XML attribute *level* of every feedback class. The default feedback level is zero. At any given time, only feedback classes belonging to a single level will be executed. Initially, all feedback classes of level zero are executed. The feedback manager continuously monitors every feedback class' valence. If all classes of level x are rated undesirable for a predefined amount of time, the manager will switch to executing classes of level $x+1$, if $x < x_{max}$. The opposite is also true. If all classes of level x are rated desirable for a predefined amount of time, the manager will switch to executing classes of level $x-1$, if $x > 0$. The time the feedback manager waits until changing the level is defined using the *progression* and *regression* options of the SSJ component.

Using this mechanic various adaptive feedback strategies can be defined. For example,

a modality progression as described in Section 4.3.2 can be achieved using the following strategy:

```
<strategy>
 <!-- level 0 -->
 <feedback type="visual" layout="table" level="0" valence="Desirable">
  <condition type="SpeechRate" event="sr" sender="ssj" from="0" to="4"/>
  <action res="ok.jpg"/>
 </feedback>

 <feedback type="visual" layout="table" level="0" valence="Undesirable">
  <condition type="SpeechRate" event="sr" sender="ssj" from="4" to="999"/>
  <action res="alert.jpg"/>
 </feedback>

 <!-- level 1 -->
 <feedback type="visual" layout="table" level="1" valence="Desirable">
  <condition type="SpeechRate" event="sr" sender="ssj" from="0" to="4"/>
  <action res="ok.jpg"/>
 </feedback>
 <feedback type="audio" level="1" valence="Desirable">
  <condition type="SpeechRate" event="sr" sender="ssj" from="0" to="4"/>
  <action res="ok.wav"/>
 </feedback>

 <feedback type="visual" layout="table" level="0" valence="Undesirable">
  <condition type="SpeechRate" event="sr" sender="ssj" from="4" to="999"/>
  <action res="alert.jpg"/>
 </feedback>
 <feedback type="audio" level="1" valence="Undesirable">
  <condition type="SpeechRate" event="sr" sender="ssj" from="4" to="999"/>
  <action res="alert.wav"/>
 </feedback>
</strategy>
```

This strategy would generate the behaviour illustrated in Figure 7.11. Initially, the manager starts at level 0. Thus, if the user's speech rate is below 4 syllables per second, just an "ok" image will be visible. Now, if the user's speech rate increases and exceeds 4 syllables per second, the user will see an "alert" image and the valence of the feedback class will switch to undesirable. If the user does not reduce their speech rate and remains in this state, after 20 seconds (as defined by the *progression* option), the manager will discontinue the execution of level 0 classes and switch to executing level 1 classes. At level 1, instead of delivering unimodal visual feedback, SSJ will deliver multimodal audio-visual feedback. More precisely, if the user's speech rate still exceeds 4 syllables per second, the "alert" image will be accompanied by an "alert" sound. Once the user adjusts their behaviour (drops below 4 syllables per second) and persists for 20 seconds, the manager will switch back to the unimodal feedback of level 0.

The large flexibility of the mechanic allows for designing numerous other adaptive feedback strategies. For instance, adaptation can also be used to increase the prominence of the feedback within the same modality.

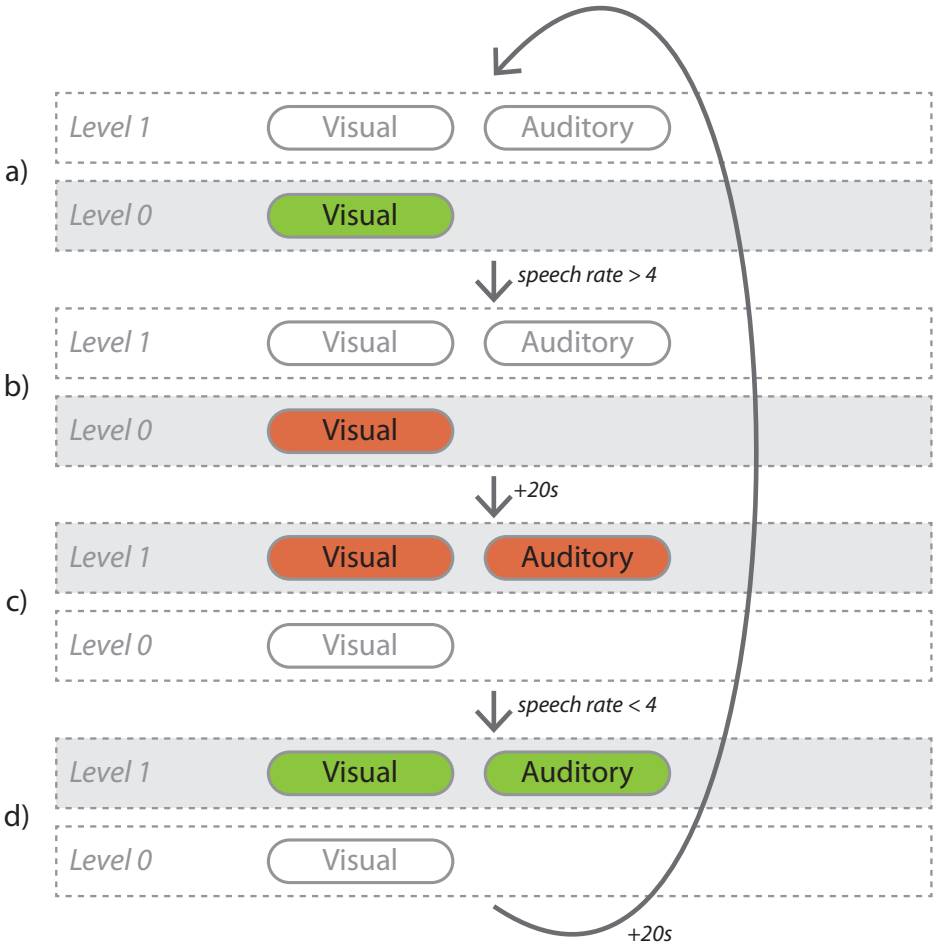

Figure 7.11: Behaviour of an adaptive feedback strategy: (a) the system starts at level 0, (b) the user's behaviour causes the visual class to switch to the undesirable state, (c) after a time period, the manager progresses to level 1, (d) the state of the feedback classes switches to desirable and after another time period the manager goes back to level 0.

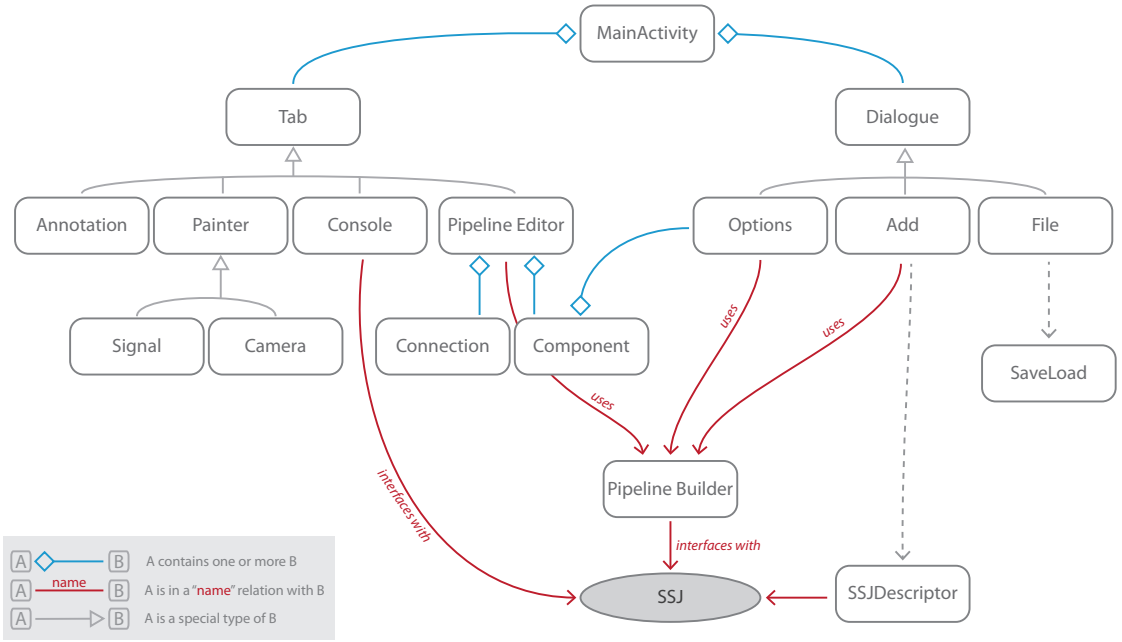

Figure 7.12: A simplified view of SSJ Creator's architecture.

7.6 The SSJ Creator GUI

In order to make SSJ accessible to persons without a technical background, a graphical user interface (GUI) called *SSJ Creator* has been implemented for the Android ecosystem. Its role is to make social signal processing and live feedback generation accessible to social scientists, students and most importantly, users of social augmentation. As discussed in Section 4.3, it is critical that the social augmentation can be tailored to the user and the environment. While automatic adaptation techniques have been proposed, it is also important that the user themself has a say in what data is recorded and how it is processed.

This section will first provide an overview of the architecture of SSJ Creator. This is followed by a description of the main application modules: pipeline manager, annotation editor and file storage system.

7.6.1 Architecture

There are two types of graphical elements in SSJ Creator: *tabs* and *dialogues* (see Figure 7.12). Tabs allow the user to inspect the SSJ pipeline from different points of view. There are four types of tabs: *annotation*, *painter*, *console* and *pipeline editor*. The annotation tab allows the user to annotate an ongoing recording. The painter tabs (*signal* and *camera*) provide a realtime view of SSJ signals. SSJ's log is displayed in the console. It provides the user with information regarding the status of the pipeline and displays runtime errors or warnings coming from SSJ. Finally, the pipeline editor allows the user to view and edit the current pipeline. The pipeline is represented in the editor using *components* and *connections*.

There are three types of dialogues in SSJ Creator: *options*, *add* and *file* dialogue. The *options* dialogue enables the user to view and edit the options of SSJ and its components. The *add* dialogue is responsible for adding new components to the pipeline whereas the *file* dialogue allows saving and loading pipelines to and from the SD card.

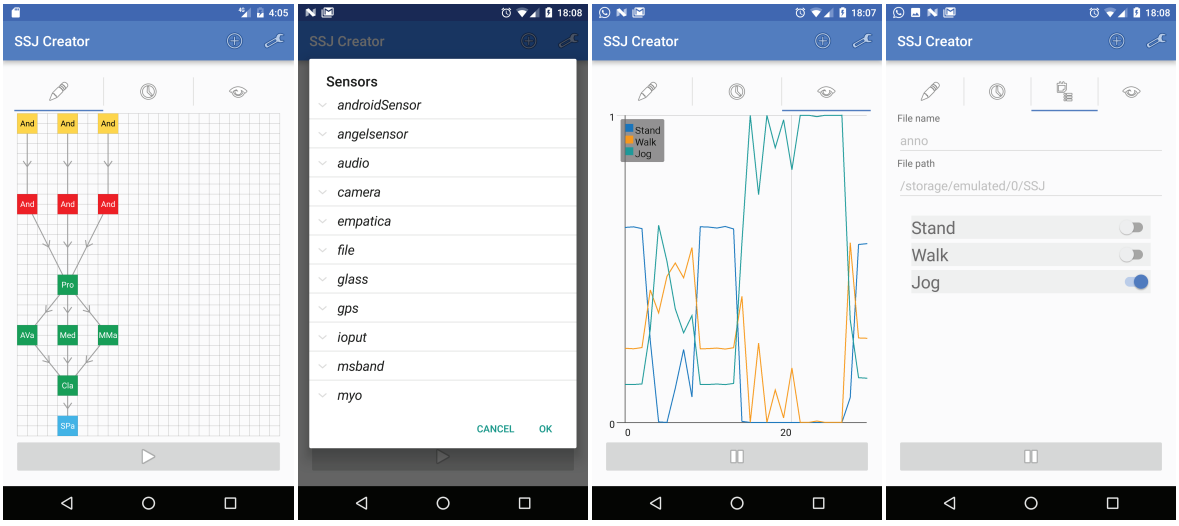

Figure 7.13: Screenshots of SSJ Creator: (a) pipeline editor, (b) adding a new component, (c) painter showing realtime feed of classification result and (d) annotation tab

The *pipeline builder* class is responsible for building and managing the SSJ pipeline. It also holds a local copy of the pipeline, which allows it to recreate the exact state of the pipeline even if SSJ shuts down incorrectly due to a runtime error. In order for SSJ Creator to know which SSJ components actually exist and populate the lists of the *add* dialogue, the entire SSJ library is parsed at system startup by the *SSJDescriptor* class. This is achieved using Android's *DexFile* interface, which allows SSJ Creator to get a complete list of all classes currently implemented in SSJ.

7.6.2 Building and Running Pipelines

SSJ Creator allows the user to build SSJ pipelines using a simple graphical user interface. Building pipelines is performed in the pipeline editor tab, which can be accessed by selecting the first tab from the left (represented by a pencil icon). This offers a view of the current pipeline configuration including components and connections between components (Figure 7.13-a). To add a new component, the user needs to access the *add* dialogue (marked by the ⊕ icon). This brings up a list of all available components grouped by type and namespace (Figure 7.13-b). If painter-type components are added to the pipeline, corresponding painter tabs are created and appended to the horizontal tab list (Figure 7.13-c).

In the pipeline view, short-tapping on a component will bring up an options dialogue where the user can alter the component's parameters. Dragging one component over another one will create a connection. If two components are connected, data will flow from the first one into the second one as discussed in Section 7.1. To remove a component, the user can simply drag it to the lower right corner of the view. Once the user finishes building a pipeline, it can be launched using the play button at the bottom of the screen. The same button also allows the pipeline to be stopped once it is running.

At runtime, SSJ will continuously send its internal log to the console tab. This allows the user to monitor the status of the pipeline and look for potential warnings or errors that could occur in response to a faulty pipeline configuration. If there are errors or the functionality is

Figure 7.14: Annotating data using the Microsoft Band 2.

not satisfactory, the pipeline can be stopped, adjusted and restarted. This allows the user to experiment with pipeline design until a satisfactory result is achieved.

7.6.3 Annotating Data

It is often the case that a recording spans multiple phases, or that it contains events which are important for the analysis. To mark such occurrences, researchers and data analysts often perform a manual post-hoc annotation of the data, during which they look at all the recordings and mark the interesting events using a specialized software. For example, in a study that contains two conditions in a within-subject design, the start and end times of each condition need to be annotated so that the data recorded during the first condition can be compared to that of the second condition. Another example is recording a training corpus for building an automatic stress classifier. Here, the stress episodes in the recordings need to be annotated to allow an automated comparison between stress and non-stress data (see example in Section 7.7).

SSJ Creator allows the user to perform realtime annotation during a recording. More specifically, once a writer-type component that stores data on the SD card has been added to the pipeline, an annotation tab will be automatically added to the tab list (Figure 7.13-d). In the annotation tab, the user can define which classes should be annotated. Once the pipeline is running, the start and end of each annotation class can be specified using a toggle button. The results of the annotation process are stored in an SSI compatible format, facilitating cross-platform analysis.

For facilitating "in the wild" data annotation by the users themselves (see Section 5.1.2), SSJ Creator also allows the annotation to be performed using the Microsoft Band 2. To achieve this, the Microsoft Band SDK is used to create a custom live tile on the Band. This gives the user access to a simple interface containing two buttons (annotation *start* and *stop*) directly on the Band (see Figure 7.14). Annotating using Microsoft Band makes the entire process of performing live annotation simpler, since the user is no longer required to take the phone out of their pocket whenever an event needs to be annotated.

7.6.4 Saving/Loading Pipelines

Building a large pipeline with multiple processing branches can be time-consuming. Moreover, it is often the case that pipelines are reused either by the same person (e.g. a recording is performed multiple times) or by a different person who wishes to replicate the recording setup.

To facilitate this, SSJ Creator allows pipelines to be stored on and loaded from the device's memory. For this, pipelines are converted into a human-readable XML format. This allows pipelines to also be edited outside of SSJ Creator using common text editors. Furthermore, since every pipeline is essentially a text file, it can be easily shared between devices and users with the help of common applications and protocols (e.g. E-mail).

7.7 Example: Providing Feedback in Response to Stress

To demonstrate the ability of SSJ to execute complex behavioural feedback loops, this section will present a tutorial on how SSJ can be used to classify stress, and then provide feedback in response to it. As a hardware platform for this task, we will use a Microsoft Band 2 and a normal Android smartphone. The Microsoft Band 2 provides access to a variety of physiological signals including heart rate, electro-dermal activity (EDA, also called GSR) and temperature. Especially the heart rate and GSR signals have been successfully used in related literature for classifying stress [Ertin et al., 2011; Gaggioli et al., 2012]. This tutorial is structured in four parts:

1. Collect and annotate data from users which are sometimes stressed, and sometimes not stressed.
2. Train a model using the collected data.
3. Build a pipeline in SSJ which uses the model to perform live classification.
4. Configure a feedback strategy to provide feedback in response to stress.

7.7.1 Data Collection

To collect the necessary data for training a model, a data recorder is needed. Using SSJ Creator, we can build a pipeline which interfaces with the Microsoft Band and stores the heart rate, GSR and temperature data locally on the device.

For this, in SSJ Creator, we first add an *MSBand* sensor using the *add* menu (see Figure 7.15-a). Following this, we add the three *MSBand* sensor channels we want to record: *HeartRateChannel*, *GSRChannel* and *SkinTempChannel*. We now connect the sensor to each individual channel by long pressing on the yellow sensor box and dragging and dropping it over each channel. Once this has been done, the pipeline should look like in Figure 7.15-b.

Finally, we need to add three *FileWriter* consumers, and connect every channel to a writer. The default behaviour of the *FileWriter* is to store the files in a timestamped folder on the SD card. If needed, this can be changed from the component options accessed by tapping on the blue *FileWriter* box. The final pipeline layout is shown in Figure 7.15-c. To start the pipeline, we simply press the "play" button. We can also save the pipeline for later use with the help of the file menu.

Once we have our data recorder in place, we need to organize a user study. The study should contain at least two tasks, one during which the participants are stressed, and one during which they are not stressed. The start of each task can be annotated using SSJ Creator's annotation tab (see Figure 7.15-d).

We recently performed such a study in association with the Charité Geriatrics research group in Berlin. A total of 10 participants (five female) aged between 61 and 86 (mean = 71.9) have been recruited. During the study, the participants first filled out a lengthy questionnaire on demographics. The content of the questionnaire was not relevant since the scope of the task was to simply simulate a non-stressful activity. Afterwards, the participants took part

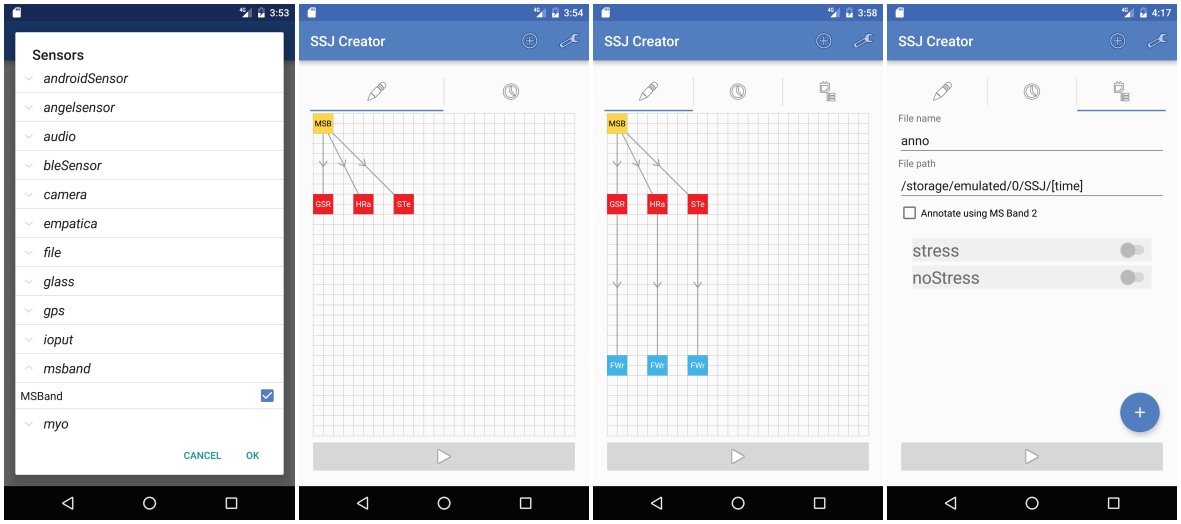

Figure 7.15: Building a data recorder using SSJ Creator: (a) adding the Microsoft Band sensor, (b) pipeline view of sensor with channels, (c) final pipeline layout (d) configured annotation tab.

in an attentional exercise that was meant to induce stress [Zimmermann and Fimm, 2004]. Using a data recorder similar to the one described above, we managed to record over 5 hours of physiological data.

7.7.2 Training a Model

Once we have collected user data, we can train a model for the automatic classification of stress. Since the initial training of machine learning models is a very resource intensive process, performing it on mobile devices is not recommended. Thus, for this tutorial, we use the Social Signal Interpretation (SSI) framework for Windows [Wagner et al., 2013]. Once an initial model has been created, online or active learning approaches could be employed to update it directly on the mobile device (see Section 10.2.5).

For training the model, we first process the raw data streams and extract features which are indicative of stress. For the purpose of this tutorial, we use the *Functionals* feature collection on every data stream. This collection includes various statistical measures such as mean, energy, standard deviation, minimum, maximum, and others. The benefit of these features is that they are also implemented in SSJ, meaning that we can easily incorporate them in the live classification pipeline as well. Finally, we use the extracted features to train a model. For this tutorial, we will use a Support Vector Machine (SVM) model since it is also compatible with SSJ. The SSI source code for the training process is provided in Appendix E.

By applying this process on the data recorded in the previously mentioned study, we managed to train an SVM which achieved a user-independent recognition rate of 83.4%. However, a more robust recognition could be achieved with more advanced processing methods. For example, before extracting the features, filters could be used to remove noise from the signals. Furthermore, additional feature extraction algorithms may also contribute to achieving better recognition rates.

7.7.3 Realtime Classification using SSJ

To perform realtime stress classification on a smartphone, the model trained in the previous section now needs to be incorporated in an SSJ pipeline. The pipeline starts by defining the sensor and sensor channels we require for data classification. Since we trained our model with heart rate, GSR and temperature, we will also use these channels in our live pipeline.

```
// Setup
Pipeline pipe = Pipeline.getInstance();

// Sensor
MSBand msBand = new MSBand();

HeartRateChannel hr = new HeartRateChannel();
pipe.addSensor(msBand, hr);

GSRChannel gsr = new GSRChannel();
pipe.addSensor(msBand, gsr);

SkinTempChannel temp = new SkinTempChannel();
pipe.addSensor(msBand, temp);
```

Now, we have to process the data exactly as we did when training the model by computing the functionals from the raw signals. To reduce recognition latency, we will configure the pipeline to call the *Functionals* transformers with a 5 second *frame* and 5 second *delta window*. This means that every transformer will receive five seconds of new data (the frame window) and 5 seconds of old data (delta window, i.e. overlap with the previous frame window). This allows the pipeline to match the 10 second time windows used for training the model and still achieve an acceptable latency of 5 seconds.

```
// Extract features from the individual data streams
Functionals gsrFeatures = new Functionals();
pipe.addTransformer(gsrFeatures, gsr, 5, 5);

Functionals hrFeatures = new Functionals();
pipe.addTransformer(hrFeatures, hr, 5, 5);

Functionals tempFeatures = new Functionals();
pipe.addTransformer(tempFeatures, temp, 5, 5);
```

All three feature streams are then fed into a *Classifier* consumer which loads the previously trained stress model to classify the live data into stress and non-stress. Finally, an *EventLogger* can be used to print out the output of the classier on the console.

```
// SVM-based classification
Classifier stress = new Classifier();
stress.options.trainerFile.set("stress_model.trainer");
stress.options.event.set("stress");
stress.options.sender.set("SSJ");
Provider[] input = new Provider[] {gsrFeatures, hrFeatures, tempFeatures};
pipe.addConsumer(stress, input, 5, 5);
EventChannel stressChannel = pipe.registerEventProvider(classifier);

EventLogger log = new EventLogger();
pipe.registerEventListener(log, stressChannel);
```

7.7.4 Live Feedback

The final step of the tutorial is to extend the live classification pipeline with a feedback manager that is able to deliver feedback to the user whenever they are stressed. To achieve this we add a *FeedbackManager* to the pipeline and register it as a listener for the *stressChannel*.

```
FeedbackManager feedback = new FeedbackManager();
feedback.options.strategy.set("strategy.xml");
pipe.registerEventListener(feedback, stressChannel);
```

In terms of feedback, we define a multimodal feedback strategy which uses both visual and audio feedback to attempt to calm the user down. The feedback is triggered every time the classifier's confidence value for stress is 0.7 or greater. To avoid the feedback becoming annoying, the audio file should be loop-friendly and the self lock should match its duration.

```
<strategy>
 <feedback type="visual">
  <condition event="stress" sender="SSJ" from="0.7" to="1.0"/>
  <action res="relax.jpg"/>
 </feedback>
 <feedback type="audio">
  <condition event="stress" sender="SSJ" from="0.7" to="1.0"/>
  <action res="relax.mp3" lockSelf="30000"/>
 </feedback>
</strategy>
```

The next code snippet starts the pipeline, and after a period, shuts it down.

```
pipe.Start();

// the pipeline is now running, this thread can wait
Thread.sleep(60 * 1000);

pipe.Stop();
```

This is it. Once the pipeline is active, it will continuously extract data from the Microsoft Band and classify it into stress and non-stress with the help of the pre-trained model. If stress is classified, audio-visual feedback will be delivered to the user. The feedback is delivered on the device running the pipeline, in our case, the smartphone. Alternatively, one could split up the pipeline between the smartphone and an HMD (e.g. Google Glass). The main pipeline performs the classification on the computationally superior smartphone and sends the results over Bluetooth to the HMD. A second pipeline is executed on the HMD. It receives the Bluetooth events and pushes them into the feedback manager for feedback generation. The strategy can be extended to also deliver tactile feedback on the Microsoft Band 2.

7.8 Summary

This chapter introduced the SSJ software framework for creating social augmentation systems. SSJ supports the behavioural feedback loop in its entirety, offering both mobile social signal processing and live feedback functionality. On the behaviour analysis side, it is able to interface with 12 sensors and offers a rich selection of processing and classification algorithms. In terms of feedback, SSJ can use four different output devices for delivering feedback spanning three modalities: audio, visual and tactile. Thanks to an efficient resource

management, SSJ is able to run on almost all Android devices including smartphones, tablets, smart glasses and smart watches.

To allow SSJ to be used by non-technical persons such as social augmentation users, a user-friendly GUI has also been implemented. It enables the design and execution of SSJ pipelines without writing a single line of code.

Thanks to its general methodology and flexible design, SSJ can also be used in non-augmentation scenarios. The mobile social signal processing tools of SSJ are state-of-the-art and can easily compete with dedicated signal processing frameworks. For example, SSJ has been successfully used as part of the Glassistant project[7] to analyse the behaviour of elderly persons, and detect when they are in need of assistance. To maximize the contribution to the research community, the entire framework is open source and freely available for download.[8]

[7] http://glassistant.de

[8] http://hcm-lab.de/ssj

8. Augmenting Public Speaking

Public speaking is a distinct type of social interaction. Some persons enjoy holding speeches, most don't. Recent estimates suggest that over 70% of the general population experiences some form of anxiety or nervousness when it comes to speaking in public [Hamilton, 2011; Richmond and McCroskey, 1998]. There is even evidence that public speaking "is the most common fear", being named more often than any other fear, including death [Dwyer and Davidson, 2012]. Public speakers need not only deliver a convincing message to their audience, but also inspire and generate enthusiasm at the same time. To achieve this, the speakers need to master their verbal and especially their nonverbal behaviour. For example, even a highly interesting message, if delivered with a slow and monotonous voice may cause boredom. On the other hand, if the speech rate is too high, people might have difficulty understanding the message at all. The key lies in the balance. This makes public speaking an ideal candidate for social augmentation. In such a scenario, a social augmentation system can help relieve some pressure by delivering direct and objective feedback on the quality of one's behaviour as well as providing instructions on how to improve it.

This chapter[1] introduces an example for a public speaking augmentation system using behavioural feedback loops. More specifically, it presents the *Logue* system. *Logue* is designed to provide in-situ and realtime feedback on a speaker's nonverbal communication unobtrusively during a public speaking scenario using a wearable head-mounted display (HMD). In a typical social augmentation context, the feedback aims to increase the users' awareness of their own nonverbal behaviour, as well as inform of the behaviour's appropriateness in the given scenario. To achieve this, social signal processing techniques are employed to analyse the speaker's performance using data from a microphone and a depth camera. Based on this analysis, feedback is generated and delivered to the user on speech rate, body energy and openness in realtime using an HMD.

Over the course of two user studies, we measured the effect of the system on the user's

[1] This chapter is an adaptation of Damian et al. [2015b].

Figure 8.1: System setup: User wearing the HMD and microphone (far plane), and a Microsoft Kinect oriented towards him (near plane).

behaviours (a) objectively using signal processing techniques as well as subjectively (b) by other persons and (c) by the users themselves with the help of surveying techniques.

8.1 System Overview

Logue consists of two main components. First, the behavioural analysis component is responsible for perceiving, processing and classifying the user's nonverbal behaviour using various sensors. The resulting analysis is sent to the feedback generation component where it is converted to visual feedback displayed on the HMD. This way, three parallel behavioural feedback loops are created, each one pertaining to a different social signal: speech rate, movement energy and posture openness.

While the system presented here is not fully mobile, being still bound to a desktop computer for running SSI and interfacing with a Microsoft Kinect, a similar functionality has been achieved on a mobile platform using the SSJ framework introduced in Chapter 7. Appendix F provides the corresponding SSJ pipeline and feedback strategy.

8.1.1 Behaviour Analysis

In the scenario of public speaking, body expressivity and vocal quality are good features for measuring the quality of a speaker's nonverbal behaviour. Since the main goal of the system is to improve the user's awareness of their own nonverbal behaviour, we explicitly avoided analysing presentation-related characteristics (e.g. total time, time-per-slide). Furthermore, using small-scale in-lab pretests, we narrowed down the problem to social signals that are technically feasible to be analysed and classified robustly in realtime. For example, while vocal clearness and loudness are good candidates, pretests showed that the analysis is very susceptible to environmental noise, the position of the microphone and non-speech sounds (throat clearing, coughing). We ended up with three social signals that allow us to provide the speaker with feedback on speech rate, body energy and openness.

To measure these three signals, we use the Microsoft Kinect depth camera and the SHURE WH20 close-talk microphone paired to a TASCAM US 322 audio interface (Figure 8.1). The

data analysis is performed using the SSI framework [Wagner et al., 2013]. To compute the speech rate in realtime, we rely on work done by de Jong and Wempe [2009] to split voiced audio segments into syllables. The number of syllables is then divided by the length of the utterance to yield the speech rate.

Body energy is measured from the tracked positions of the user's hands in accordance with related literature [Baur et al., 2013b; Caridakis et al., 2006]. More precisely, we use the position of the wrist joints (as these are more robustly tracked by the Kinect than the hands) relative to the neck joint to compute the spatial displacement over the course of 5 seconds. The displacement is normalized using the arm span of the user to make the measurement user independent.

Openness is computed similarly. The difference is that instead of measuring the spatial displacement, we compute the Euclidean distance between the hands. However, unlike spatial extent [Baur et al., 2013b], openness can also be negative if the arms cross.

To reduce the frequency of feedback events, before forwarding the analysis results to the feedback component, we apply a moving average filter to each signal with a window of five utterances (speech rate), 50 seconds (energy) and 30 seconds (openness). The goal here is to analyse the behaviour of the user as a whole rather than focus on fluctuations during the interaction. After this final processing step, the results of the behaviour analysis are forwarded asynchronously to the feedback generation module.

8.1.2 Feedback Delivery

The results of the behaviour analysis are converted into symbolic visual feedback by the feedback generation module, which is based on an early version of SSJ's feedback manager (see Section 7.5). More specifically, each behavioural feedback loop is mapped to a *functional icon* rendered at a fixed position on the HMD. Since the focus of the system is to increase the users' awareness of their own behaviour, the functional icons are designed to deliver appraisive feedback. More precisely, each functional icon can express three intensities of the current behaviour as well as provide information on whether the behaviour's current state is appropriate or not.

For each social signal, two thresholds (a lower and an upper threshold) determine how the behaviour state is mapped onto the intensities of the functional icons (low, medium or high). The thresholds also generate an appropriateness corridor, with behaviours within this corridor being marked appropriate, and those outside inappropriate. These thresholds were configured beforehand in a small-scale study, during which three users purposely performed well and bad for each of the social signal. The resulting data was averaged across all participants and used to compute the two thresholds for each behaviour. While these thresholds are meant to describe normal speaking behaviour, one could also choose other thresholds for very specific situations, e.g. to lead a presenter to perform in a highly energetic way using a loud voice.

Each functional icon is composed out of two symbols positioned one on top of the other (see Figure 8.2). The upper symbol reflects the behaviour analysis and classifies the intensity of the behaviour into three classes (low, medium, high). The appropriateness of the current behaviour is represented using the second, smaller symbol. The main reason for separating the appropriateness from the intensity was to ensure correct and fast recognition on see-through displays by having both channels encoded in both shape and colour. Such displays are notorious for causing colour perception difficulties in uncontrolled environments [Gabbard et al., 2010] (also see evaluation in Section 6.1.1).

Figure 8.2: Illustration of user's field of view showing the visual feedback in the upper right corner.

The visual feedback is a high duration feedback, being displayed persistently on the HMD. The persistent display of the feedback makes it become a familiar perceptual object to the user's sensory system, reducing the probability of an involuntary attentional capture (see Section 2.2.2). This, combined with the position of the icons in the user's field of view, namely in the upper right corner, reduces the overall prominence of the feedback and with it, the probability of disrupting the primary task.

The rendering of the visual feedback is handled by a Vuzix STAR 1200 optical see-through HMD (depicted in Figure 8.1). It features an 1280x720 resolution spanned over two see-through displays, which offer a 23 inch diagonal field of view. The system has also been successfully deployed on different HMDs, such as the Google Glass.

Icons

For our social augmentation task, it is important that the feedback icons do not distract the users from their primary task (R2 – see Section 4.1), i.e. the social interaction itself, while still generating awareness and providing guidance (R1 and R3). Thus, we conducted a pre-study to help us select the most appropriate icons for each behaviour out of an initial set of 33 icons (Figure 8.3). In an effort to be understandable at a glance when viewed on a see-through HMD, and to demand only minimal attention for processing, all icons have been designed to provide a symbolic representation of information. The icon set spans our three feedback loops, i.e. *speech rate*, *energy* and *openness*, and six intensity themes inspired by [Bertin, 1983; Szczerba et al., 2012] (shape, area, orientation, size, composition and quantity). The set also includes six additional icons covering the two appropriateness classes *positive* and *negative*.

For the user study, we split the icons into groups, each pertaining to a single behaviour (as shown in Figure 8.3), and asked 25 students (4 female, 21 male; mean age of 22.4) from our university to rate on a scale from 1 (worst) to 7 (best) how representative of the behaviour each icon is. The best rated icon groups (highlighted in Figure 8.3 with a dark grey background) for each behaviour have been implemented in the system.

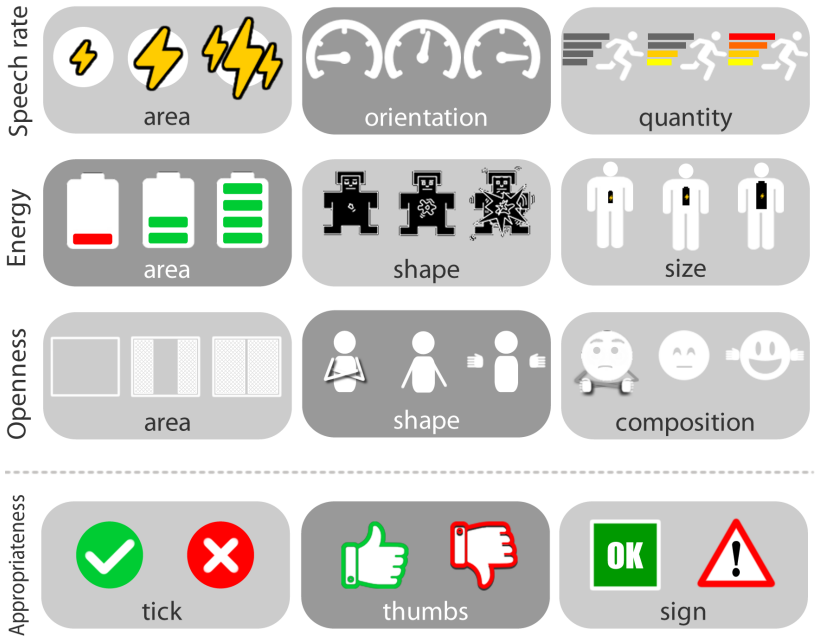

Figure 8.3: Initial icon set categorized by behaviour and theme. Highlighted icon groups have been found to be most suited for delivering feedback on their particular behaviour.

8.2 Evaluation

To ascertain the effectiveness of a public speaking augmentation, we [Damian et al., 2015b] conducted two user studies with the help of the system described before. For the first study, a mixed group of participants were recruited from the university to use the system in a controlled environment. They were given the task of presenting an "elevator pitch" speech to a small, targeted audience. The goal of their talks was to convince the targeted audience, which played the role of potential investors, to invest in their projects. The study aimed to quantitatively determine the impact of the augmentation on the participants' presentation performance.

For the second study, we asked senior PhD students to test the system during a presentation that had to be given to peers and supervising professors at an annually organized PhD workshop. The goal of this study was to understand how the augmentation is perceived in a real setting – i.e. one that was not staged for the purpose of the study.

8.2.1 Study One: Quantitative Evaluation

The first study focuses on the collection of questionnaire data and measurement of the social signals of the participants. This allowed us to compute both subjective and objective measurements of the participants' performance.

Participants and Apparatus

We recruited a total of 15 computer science undergrad and recently graduated students (13 male and 2 female) with an average age of 26.13 (henceforth referred to as P1, $\cdots$ P15). On a 7-point Likert scale (1 = worst, 7 = very good), the participants rated themselves 3.33 for frequency of holding presentations and 4.07 on how skilful they think they are. Two employees of the institute (aged 28 and 35) were recruited to act as observers during the

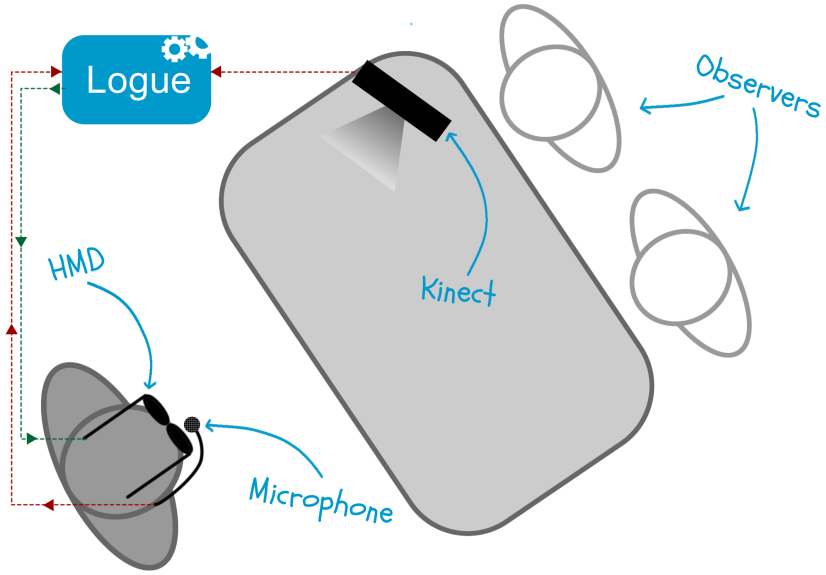

Figure 8.4: Evaluation setup showing the participant facing two observers while wearing the HMD and microphone. The Microsoft Kinect was positioned on the conference table between the participant and observers, and was oriented towards the participant.

whole study.

The study was held in a typical conference room with the participant standing at the front of the room facing two observers at a distance of 3 m. The observers were instructed to pay attention to the nonverbal behaviour of the participants. Each participant wore the Vuzix HMD and a head-worn microphone. The Kinect was positioned in a way that did not obstruct the observers' view of the participant (Figure 8.4).

Task and Procedure

The participants were asked to perform two public speeches for two conditions (control and experimental) in a within subject design, i.e. one speech for each condition. In the control condition (CC) the users wore the system, but the feedback visualisation was deactivated. This condition provides a baseline for quantitative comparison as no feedback was presented to the participants. In the experimental condition (EC) the participants received feedback on their nonverbal behaviour using the proposed system. The observers were blind to how the conditions were assigned to the participants. To minimize learning effect, the two sessions were scheduled to be roughly two weeks apart and the order of the conditions was randomized. The topics of the speeches were various software projects on which the participants worked in the past half year. Each talk was expected to last approximately five minutes and no supportive materials (e.g. slides) were allowed. The participants were told to act as if the observers were potential investors, who might be interested in investing in their projects.

Prior to each session, participants filled out a questionnaire about their experience with regard to public speeches.[2] The participants were then given instructions on how to use the system and the feedback mechanism was explained in detail. Emphasis was placed on making sure that each student had fully understood the correlation between their behaviour and the feedback icons. Afterwards, the participants were given five minutes to familiarize themselves

[2] http://hcm-lab.de/downloads/chi15

with the system and ask any questions to the experimenters.

After each session, both the participant and the observers filled out a second questionnaire meant to elicit data regarding the participant's performance as a public speaker and perceived user experience.[2] Lastly, a semi-structured interview with the user to gather general feedback about the system was conducted.

Results

We recorded video, audio, depth and social signal data from a total of 30 sessions in addition to the questionnaire data. The average length of each session was 4 minutes and 18 seconds. The audio recording from one participant had to be excluded from the analysis due to a medical condition that caused frequent throat clearings and interfered with the audio analysis. Similarly, the recording of one participant was omitted from the energy and openness analysis due to skeleton tracking problems.

(A) Impact on objective measurements.

We processed the recorded data for each feedback class by computing how many of the participants' vocal segments ("utterances") measured speech rates outside of the thresholds. We normalized this value by the total amount of utterances of the session and averaged it over all participants to get a measure of the speech rate inappropriateness for each condition. Energy and openness was processed analogously . However, since the energy and openness are computed continuously (unlike the speech rate), instead of the number of utterances, we used the normalized duration in seconds of the time spent outside of the thresholds.

The resulting values can be seen in Figure 8.5. A One-Way Repeated Measure MANOVA revealed by trend a multivariate effect of social augmentation on the amount of inappropriate behaviour for speech rate, energy and openness, $F(3,11) = 3.215$, $p = .065$. When looking at the feedback classes individually, a Wilcoxon signed rank test showed significant differences between the conditions for speech rate with $p = 0.033$, $Z = -2.134$. There were no significant effects for energy and openness though. As Figure 8.5 illustrates, we measured increased standard errors of the mean for all feedback classes. These were caused by the participants who never crossed the thresholds and thus yielded zero amount of inappropriate behaviour.

When taking a closer look at the data, we can observe how the participants reacted to the feedback. Figure 8.6 shows P13's openness over time and how it was affected by the behavioural feedback (symbolized by the vertical lines). As can be seen in the figure, P13 immediately reacted to the system's feedback by performing more open gestures.

(B) Impact on observer's perception.

The observers rated the participants as significantly more open ($F(1,14) = 3.333$, $p = 0.045$) when the social augmentation mode was activated than when it was not. This suggests that the social augmentation mode had an observable impact on the participants' behaviour. However, we did not find any significant differences for the other dimensions of the observers' questionnaire.

(C) Impact on self-perception.

T-tests for one sample revealed that the participants' subjective ratings of the social augmentation mode were significantly above the neutral value of 4.0 (Figure 8.7). The participants thought the feedback was correct ($t(13) = 6.77$, $p < .0005$), helpful ($t(14) = 7.25$, $p < .0005$) and not confusing ($t(14) = -5.332$, $p < .0005$). The feedback also gave them a sense of se-

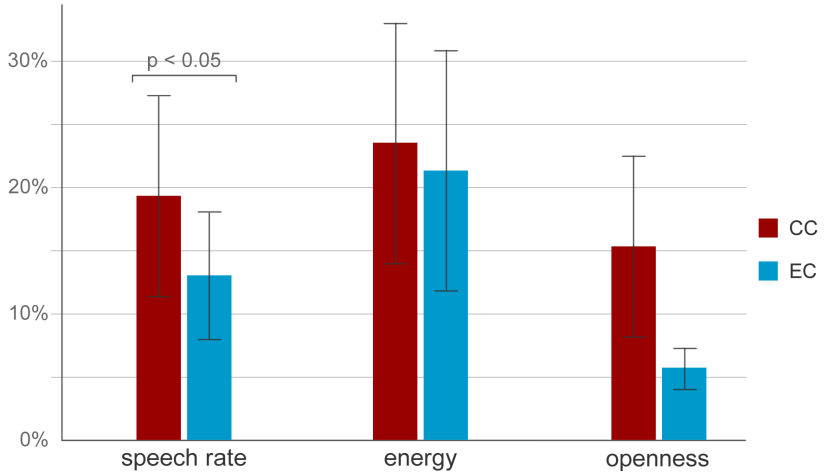

Figure 8.5: Percentage of inappropriate behaviour (y-axis) for each feedback class across conditions (control vs. experimental). Lower values are better.

curity ($t(14) = 4.75$, $p < .0005$) and they did not try to ignore the feedback ($t(14) = -6.094$, $p < .0005$). A comparison between the conditions yielded no evidence that the participants felt distracted by the social augmentation mode or that the social augmentation mode increased the difficulty of the main task.

8.2.2 Study Two: Qualitative evaluation in a real setting

Following our first study, where we made use of an enacted scenario, we conducted a second study in order to test the system in a real presentation setting. To this end, we recruited speakers from an annual doctoral workshop who volunteered to make use of the system during their presentation. The workshop gives computer science PhD students the opportunity to present the current state of their work to their peers and supervisors. The experimental setting did not only expose the participants to more realistic conditions, but the quality of the presentation also had an impact on the participants' standing within the laboratory. Apart from the fact that some people made use of the system during their presentation, the workshop was held similarly as the years before.

Participants and Apparatus

For the second study, three senior computer science PhD students (P16, P17 and P18) were recruited. Similar to the other workshop participants, they had to give a prepared presentation on their PhD topics. The talks took place in a seminar room, with the participant standing at the front of the room and facing the audience, which consisted of 13 peers and two supervising professors. All speakers made use of slides to accompany their presentation. The presentations lasted about 30 minutes and typically included ten minutes of discussion.

Task and Procedure

After the talk and the scientific discussion, an open discussion on the style of presentation followed. The audience was asked questions regarding the perceived quality of the talk as well as whether they felt the proposed system had influenced the quality of the presentation in a positive or negative manner. Later, we conducted a semi-structured interview with each participant to elicit their impression of the system.

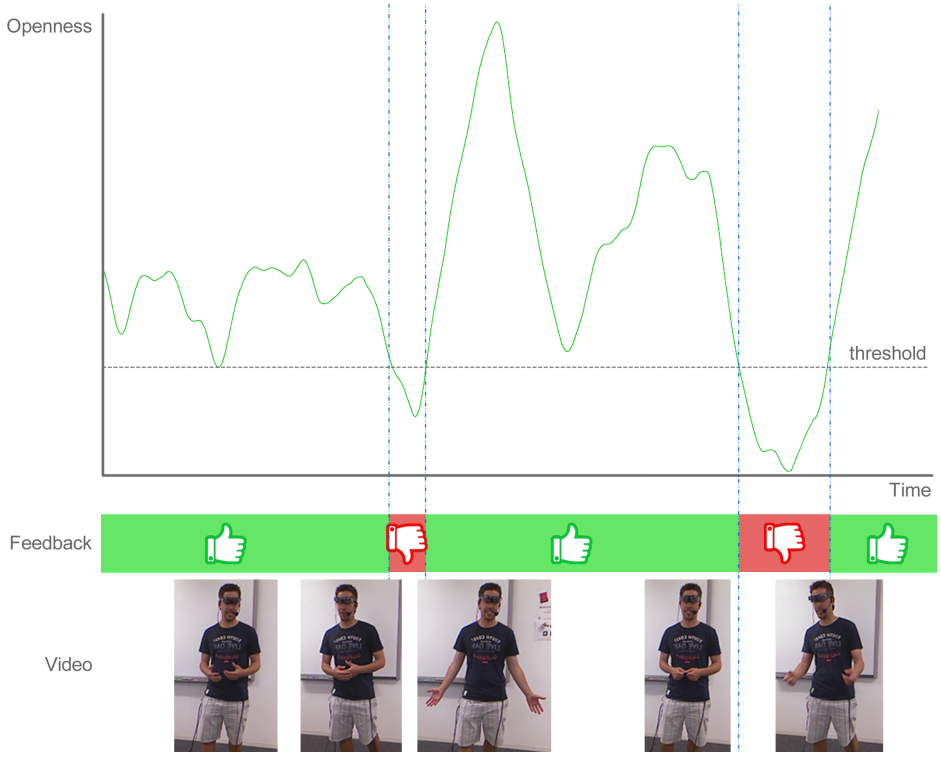

Figure 8.6: Example of participant's openness over the course of a session in the experimental condition.

Results

We highlight three outcomes of the semi-structured interviews regarding the augmentation's impact on behaviour, its level of distraction and its usefulness.

Did you adapt your behaviour?

Speaker 1 (P16) stated "One time during the presentation I felt I was talking too fast and then I remembered I had this thing on my nose so I looked at the feedback and this was actually the case. I then tried to talk slower." We then asked the audience what they thought about the speech rate. One observer said "[the speech rate] was not actually disturbing, but it would have helped if he had talked slower," indicating that there were indeed issues with the speaker's speech rate. P17, who told us prior to the study that he is aware he talks really fast, explained: "I was surprised that the speech rate did not become red sooner [...] once I saw the feedback that I was talking too fast, I tried to adapt." Hence, our system had an effect on their presentation and they tried to adapt their behaviour as suggested by the system.

Was the system distracting?

All participants complained about the bulkiness of the HMD. Despite this, P16 added "it is a very interesting concept, most of the time I did not perceive the system, only when I consciously looked at the feedback. It would be interesting to try it out for a longer period, for example to use it when teaching a class ... but with the Google Glass." P17 also said "the system was unobtrusive [...] I consciously looked at the feedback from time to time." We take this as a positive sign that the system does not pose an unacceptable distraction for the speakers.

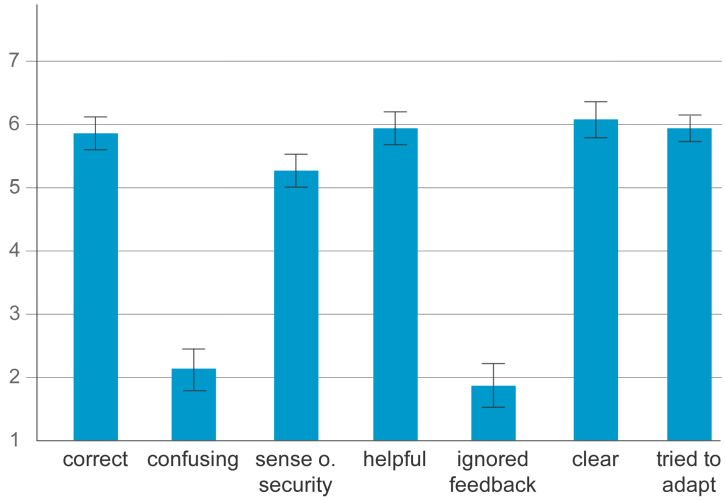

Figure 8.7: Results of the user experience questionnaire showing means on a 7-point Likert scale (1 = worst, 7 = very good). Two items (*confusing* and *ignored feedback*) are reverse-scored.

Would you use it?

When asked whether they would use the system, all were positive. P16 said: "I would use it during real presentations or while teaching a class. Or to train for a presentation;" P17 was more reserved and said "If you used it regularly you would get a feeling for what's good or bad so that you might not need it any more after a while." These statements are encouraging, suggesting that the augmentation provides sufficient value to warrant its regular use, for example during lectures.

8.2.3 Discussion

Effectiveness in Altering Behaviour

The results from the first study suggest that the system had a beneficial impact on the behaviour of the participants. For all three feedback classes, we were able to measure improvements over the control condition in terms of amount of appropriate behaviour as measured by the system, with significant effects for the speech rate. Thus, the proposed augmentation clearly respects the third and fourth requirement of augmenting social interaction (see Section 4.1), which state that the feedback is appropriate for facilitating a behavioural change and that this change is beneficial to the interaction.

To get a better understanding of how the system impacted the behaviour of a speaker, we refer to Figure 8.6. It shows how P13 adapted his openness after a visual feedback event, but also how it slowly degraded over time until another correction was needed. This effect further strengthens the potential of the proposed approach, as it acts similar to a reinforcement technique, repeatedly reminding the users to adapt their behaviours. This effect was found among multiple participants and is backed up by statements from the semi-structured interviews from both studies. For instance, P1 stated "I think the system helped me a lot as it reminded me to talk slower. This is a problem many people pointed out to me after I hold a presentation and it felt good to get the feedback during the talk."

In some cases we found evidence of a longer term effect on the participants. During his second session (CC), P10 used very broad gestures when describing his project's composition.

When confronted, he admitted to remembering the feedback from the first session (EC) despite them being two weeks apart. While this effect may have had a negative impact on the actual goal of the first user study, it is nevertheless encouraging for the concept of social augmentation.

Need for personalization

It is interesting to note that some participants did not cross the appropriateness thresholds at all. Furthermore, participants seemed to react differently to the feedback. Some adapted the behaviour instantly, others gradually and a few participants ignored it. This can be explained by a heterogeneous distribution of presentation skill among the participants. If we classify the participants into experts and novices based on their answers from the demographic questionnaire, we notice some interesting trends. Novice users improved their speech rate more than experts (mean decrease in amount of inappropriate behaviour $\Delta = 9.3\%$ versus $\Delta = 0.9\%$). This shows that experts are already speaking at an optimal speech rate. Remarkably, when looking at the openness dimension, the effect is reversed. The experts managed to improve more ($\Delta = 31.5\%$) than the novices ($\Delta = 1.2\%$). A possible explanation for this is that the participants who rated themselves bad at public speaking (i.e. the novices) appeared to be more introverted and thus had difficulties performing open gestures even when given feedback to do so. For example, during the post-hoc interview, P1 said "I did not know how to correct my behaviour. I tried to perform more gestures, but I found it very difficult to incorporate these into my performance." A similar statement came from P15: "I knew I had to move more, but I was afraid that simply moving my hands would look weird."

We argue that these differences denote the importance for an adaptable system. Thus, in order to fulfil the third requirement (feedback is appropriate for triggering behavioural change), it is important that the system provides not only scenario-specific customization but also user-specific individualization. The proposed adaptation mechanisms introduced in Section 4.3 could be a solution here. This is backed up by another observation during the post-interaction interviews: Despite all participants receiving equally detailed feedback, some participants stated that they were overwhelmed by the feedback whereas others asked if we can increase the level of detail on the behaviour analysis. We received similar feedback from the second study, with P16 stating: "It would be good to know how far above the threshold I am." P17 also noticed the need for customization when saying that in its current state, the system "is good for newbies" but "probably not accurate enough for professionals."

Participants also stated that they can envision themselves using the system as preparation for stressful social situations such as public speeches. In such training environments, an increased level of detail for the peripheral feedback would be feasible, as the primary task is no longer the actual interaction, but the training itself.

Primary Task Disruptiveness

To ascertain the degree to which the proposed system fulfils the second requirement of social interaction (feedback does not disrupt interaction), we asked the participants of the second study whether the system distracted them from giving the talk. The participants assured us that the impact was minimal. P15 stated "It felt distracting at first, but then I noticed that I can look through the displayed feedback. Once this was clear, I only glanced at the feedback to see whether anything has changed." Other participants, including the senior PhD students from the second study, made similar statements, saying that they were checking on the behaviour from time to time during speech or thinking pauses to see if something had changed. These

results support our initial design choices, in particular the use of persistent visual feedback. We argue that in this manner, the participant is able to decide when to access the information, thus minimizing the impact on the primary task. A more in depth analysis of this aspect is provided in Chapter 9.

Sense of security

The semi-structured interviews also revealed an increased sense of security for the participants in the experimental condition. This was even the case for participants who behaved appropriately the whole time. P6 expressed: "I would look at the icons from time to time and seeing them green made me feel better about my performance." The participants who did the control condition after the experimental condition even stated that they "missed" the system as they were unsure of the appropriateness of their behaviour. We received similar statements from the three senior PhD students in the second study, who can be regarded as more experienced speakers. Both P17 and P18 mentioned an increased sense of security, with P18 saying "It was a good feeling seeing everything green ... it's like applause, or as if someone looks at you and nods. However, the green lasts longer than a nod [laughs]." P17 was more cautious: "It could help with feeling more secure [...] if you are untrained and get cold feet."

Impact on External Perception

At the end of the second study we also asked the audience after every presentation for their thoughts and feelings on the system. Overall, they mentioned that the system appeared odd to them in the beginning because of the bulkiness of the HMD. However, they soon got used to it and it did not distract them from the actual presentation. This is encouraging and suggests that the system did fulfil R5 (augmentation system has minimal impact on interaction and interlocutors). Moreover, it is very likely that a lighter HMD would increase the acceptability in the eyes of the audience. No privacy-related concerns were raised by the audience or the speakers themselves. However, this is most likely due to the study setup and the technical affinity of the participants. Nevertheless, we argue that as long as the system only analyses the user's own behaviour, the augmentation does not violate the privacy of surrounding persons, and thus respects the final requirement.

Limitations

A major limitation of this early implementation of augmented social interaction was the HMD, as almost all participants complained about its weight and size. However, these issues would be mitigated by using a lighter HMD, such as the Google Glass.

Although the system had a positive effect on the participants' behaviour as measured by the system, no major effects were noticeable on the perception of the observers. This effect is worrisome from the point of view of R4 – i.e. the behavioural change brought by the augmentation is advantageous for the position of the user in the social interaction. One possible explanation for this is that the quality of a presentation is a very subjective measure, making the search for an objective definition of a perfect presentation futile. Most of us have witnessed good talks, but which were nothing alike. Considering this, it is possible that for our observers, different feedback classes and thresholds might have had a larger effect on their perception of the talks. The fact that we only had two observers (the objective sensor measurements were the study's focus) is also likely to have contributed. However, this problem is shared by other researchers as well. For example, Tanveer et al. [2015] also report

poor agreement on subjective assessment of the quality of public speeches despite using a larger number of observers.

8.3 Summary

This chapter introduced one concrete example of a social augmentation system for the scenario of public speaking. Using visual appraisive feedback, the proposed system increases the user's awareness of their own behaviour. More precisely, it provides the user with behavioural feedback on two levels. First, it informs the user of the current state of their speech rate (How fast am I talking?), body energy (How much do I gesticulate?) and openness (How open is my posture?). Secondly, the system indicates the quality of each of these three social signals in relation to the public speaking context. For example, a high speech rate would be marked as inappropriate as it may impact the listeners' ability to follow the talk.

We evaluated the system not only in a staged, but also in a real presentation setting to see how it impacts the behaviour of the user. Both studies yielded promising results, with the system fulfilling all six requirements of social augmentation. More precisely, objective and subjective results suggests that the augmentation had a positive effect the participants' performance (R1, R3, R4 – see Section 4.1) while not distracting from the main task (R2). Moreover, statements from post-hoc interviews reveal that despite a rather bulky HMD, the users of the augmentation as well as the members of the audience quickly got accustomed to the system, and that it did not significantly interfere with the interaction itself (R5). Finally, no privacy-related complaints have been put forwards by the participants or the bystanders (R6).

It is to be expected that an updated version of this system using the SSJ framework introduced in Section 7 would yield better results, especially in terms of social interaction disruption (R5). It would also allow the augmentation to be less restrictive thanks to the use of fully mobile devices and more tolerant to variations in user personality and skill level thanks to the automatic feedback adaptation mechanics.

Nevertheless, the evaluation left some open questions. First, the studies did not objectively measure the distraction effect of the augmentation. Moreover, it is also not clear what role the feedback modality played in the overall augmentation, or whether the results would carry over to a different scenario. To address these points, the following chapter presents a follow-up study which compares between three different feedback modalities for augmenting group discussions, and also introduces measures for objectively ascertaining the disruptiveness of the augmentation.

9. Augmenting Group Discussions

The previous chapter demonstrated the benefit of social augmentation while speaking in public. While this is a fitting example of a scenario with which a great amount of persons struggle with and thus would benefit from assistance, it is a very specialized case of social interaction. Conversation in a public speaking scenario is mostly one sided and exchanges between speaker and audience are limited. The question arises how well social augmentation would perform during more traditional interactions involving close distance face-to-face encounters. Such social interactions have been mostly unaffected by the introduction of ever more diverse and ubiquitous technologies and an ever more intrusive media channel: We still look each other in the eyes and communicate using gestures and spoken language. Diverting from this archaic set of rules (e.g. texting while talking) is usually considered rude and impolite.

This chapter[1] explores this issue. More specifically, we conducted a user study with 54 participants which were engaged in group discussions. During the discussion, each participant was supported by a behavioural feedback loop which helped the user with controlling their speaking time. The augmentations' main aim was to improve the balance and thus overall quality of the group discussion. The behavioural feedback loop was deliberately chosen to be simple in an effort to reduce the impact of personal variations in feedback interpretation and behaviour adaptation. Thus, we were able to more accurately investigate the effectiveness and disruptiveness of the augmentation while also comparing between four different feedback methods: auditory, tactile, visual (head-mounted display, HMD) and visual (remote display, tablet).

[1] This chapter is an adaptation of Damian et al. [2015a].

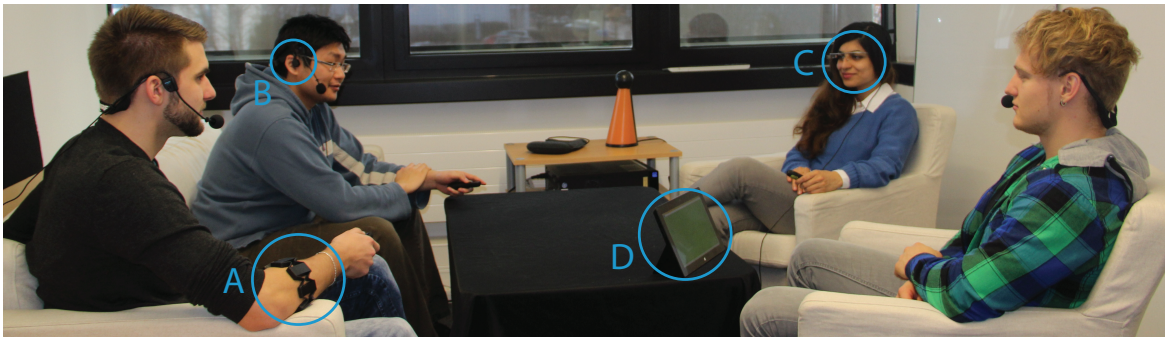

Figure 9.1: Setup of user study showing four participants, each wearing an output device: Myo armband (A), Aftershokz Bluez 2S bone conduction headphones (B), Google Glass (C), Microsoft Surface 2 Pro (D).

9.1 System Overview

The social augmentation system has been implemented using SSI [Wagner et al., 2013] and a predecessor of SSJ's feedback manager. Since the study was conducted in a laboratory, mobility was not a main concern. However, as was the case for the public speaking augmentation (Chapter 8), the functionality of the augmentation system used for this study can also be achieved with the software framework introduced in Chapter 7 (see Appendix G). This would have the benefit of making the augmentation fully mobile and capable of "in the wild" deployment.

9.1.1 Behaviour Analysis

The behaviour analysis for the augmentation was handled by a social signal processing pipeline. The pipeline extracted audio data from head worn microphones and classified it in voice and non-voice using a signal-to-noise ratio-based voice activity detector. Since the aim of the augmentation was to balance the contribution of every user to the discussion, a fixed speaking duration of 120 seconds was imposed for every user. Using the results from the voice activity detector, the remaining speaking time was continuously updated for every participant.

9.1.2 Feedback Delivery

For feedback delivery, we used four feedback methods (see Figure 9.1): tactile feedback using the Myo Armband, auditory feedback using the Aftershokz Bluez 2S bone conduction headphones[2], head-mounted visual feedback using the Google Glass and remote visual feedback using the Microsoft Surface 2 Pro[3].

The aim of the feedback was to inform the user of how much time they still had left for talking. To reduce variability, feedback was delivered at the same intervals for every device. More precisely, feedback was delivered once the user reached 75%, 50%, 25% and 10% remaining speaking time. In the case of the auditory and the tactile devices, the feedback had a short duration, i.e. it was only delivered at the before-mentioned intervals. For the visual devices, the feedback was provided persistently, however, it was only updated at the specified

[2] https://aftershokz.com/products/bluez-2s
[3] https://www.microsoft.com/surface/en-us/devices/surface-pro-2

intervals.

The tactile feedback was generated by the Myo armband and consisted of various vibration patterns delivered to the user's forearm. More precisely, one short vibration was delivered at 75% remaining speaking time, two short vibrations at 50%, three short vibrations at 25% and one long vibration at 10%. If the user passed 0%, the Myo would deliver one long vibration every time the user started an utterance.

To provide auditory feedback, we used a pair of bone conductance headphones. The largest advantage of these headphones is their ability to propagate sound waves through the skull without disrupting the normal hearing process, allowing the user to participate normally in the conversation. The auditory feedback was verbal for the 75%, 50%, 25% and 10% marks and consisted of a simple read-out of the remaining speaking time. Once the remaining speaking time was close to zero, the verbal feedback is replaced by a beeping sound which continued after the time has elapsed if the user attempted to speak.

For both visual devices, the feedback was delivered using simple colour and text coding. More precisely, a feedback event consisted of the remaining time displayed in text form in the middle of the screen. The text was displayed on a coloured background which varied from green to red on a YUV spectrum according to the remaining speaking time.

9.2 Evaluation

The main goal of the study was to investigate the impact of the augmentation on the social interaction using both objective paralinguistic statistics and subjective questionnaire-based measurements across four different feedback modalities: auditory, tactile, visual (HMD) and visual (tablet). To this end, we formulate three research questions.

Research Question 1 Does the augmentation improve the social interaction? How do the feedback methods compare?

The first research question concerns three distinct requirements of augmenting social interactions (see Section 4.1). More specifically, in order for the augmentation to improve the social interaction, the delivered feedback must be perceivable by the user (R1) and appropriate for facilitating a behavioural change (R3) which, once it occurs, has a positive effect on the social interaction (R4).

Research Question 2 Does the augmentation disturb the user? How do the feedback methods compare?

The second question of the study addresses the impact of the augmentation on the user. Thus, it is directly related to the second requirement of social augmentation: The social augmentation has a minimal impact on the attention level dedicated to the primary task.

Research Question 3 Does the augmentation disturb the other interlocutors? How do the feedback methods compare?

Finally, Q3 measures the impact of the augmentation on the persons the user is interacting with. Thus, it investigates whether the implemented system is fulfilling the fifth requirement of social augmentation.

9.2.1 Procedure

The user study featured a 4x2 mixed design. Table 9.2 summarizes the study design. Each evaluation session consisted of two group discussions between four participants, held on the same day back to back. During both discussions, each participant was given one of four output devices. Each participant from one group was given a different device which was chosen by the experimenter prior to the session by accounting for sensory-impairments (e.g. colour-blind persons were not given visual feedback devices).

During the first discussion, i.e. the control condition, the augmentation was inactive. The feedback loops were activated before the start of the second group discussion, i.e. the experimental condition. However, to avoid any bias, the participants were told that the augmentation was active either in the first or the second discussion, based on chance. The group discussions were mostly non-moderated. The experimenter intervened only at the beginning of each discussion round, to propose a conversation topic (chosen based on the interests of the participants), as well as in the case the discussion ended prematurely, in which case the experimenter would propose a new topic. Each group discussion was designed to last 10 minutes. If the duration of the discussion exceeded 10 minutes, the experimenter would again intervene and stop it.

During a discussion round, each participant had three tasks. First, they was asked to participate in the discussion normally while attempting to keep it balanced, i.e. all participants should talk for roughly two minutes. Secondly, they was instructed to pay attention to the feedback devices, which help maintain this balance by giving feedback on how much time there is left to talk. Thirdly, similarly to the study design of Ofek et al. [2013], each participant was given a wireless clicker which they had to press every time they noticed one of their peers received feedback while talking. From a psychological point of view, the discussion represents the primary task. The secondary task is either the feedback task (when the participant was talking) or the clicker task (when not talking). To enforce this configuration, feedback was only delivered when the participant was talking and the clicker was only allowed to be used when not talking. Prior to the first discussion round, each participant received an extensive introduction on the study tasks and the function of the output device.

9.2.2 Measures

After each session, the participants filled out a questionnaire with items on the perceived effectiveness of their own feedback method as well the method of the others. All items consisted of a 7-point Likert scale where 1 was labelled as “I do not agree” and 7 as “I agree fully.” In total, we analysed 13 items, with 4-5 items for each research question (ix-y = item y of Qx). Table 9.1 list all the items of the questionnaire.

In addition to the questionnaires, video and audio data has been recorded for each session. For this, a small webcam was installed in the room and each participant wore a close-talk microphone. This data allowed the post-hoc analysis and comparison of the participant’s behaviour. We decided against installing any more sensors to avoid artificially altering the behaviour of the participants and keep the interaction as natural as possible.

9.2.3 Participants

A total of 54 persons (mean age = 24.6, 40 male and 14 female), split over 14 sessions, participated in the user study. Due to last-minute drop-outs, two sessions had to be carried out with only 3 participants. The gender distribution across sessions was as follows (M =

i1-1	The device supported me during the discussion
i1-2	The feedback gave me a better perception of my speaking time
i1-3	The discussion improved because of my feedback
i1-4	Because of the feedback I was able to better control my speaking time
i2-1	The feedback disturbed my conversation flow
i2-2	I would have preferred to talk without the feedback
i2-3	Regardless of the type of feedback, the device itself disturbed me
i2-4	I would have preferred to talk without the feedback device
i3-1	I think it was observable on my behaviour when I received feedback
i3-2	I think others noticed when I received feedback
i3-3	I think others were disturbed by me receiving feedback
i3-4	The feedback of the other person's [HMD/Myo/Headphones/Tablet] disturbed me
i3-5	The mere presence of the other person's [HMD/Myo/Headphones/Tablet] disturbed me

Table 9.1: Post-Session experience questionnaire with 13 items (4-5 items for each research question).

		between-subject			
		Tactile	Audio	Visual (HMD)	Visual (Tablet)
within-subject	CC	13	13	14	14
	EC	13	13	14	14

Table 9.2: Participant distribution across conditions and devices.

male, F = female): MMMM = 4, MMMF = 7, MMFF = 1, MFF = 2. The distribution of participants over the conditions is detailed in Table 9.2. Most participants were students of either computer science, physics, mathematics, law or economics. Users were compensated with snacks and a gift card raffle.

The participants were recruited two to three weeks prior to the evaluation sessions from various sources and, in most cases, had no prior connections between each other. The recruitment consisted of an online questionnaire which included items on demographics, session scheduling and interests. The data gathered from the recruitment questionnaire was used to create balanced discussion groups in terms of interests and gender.

9.2.4 Results

Objective Measurements

After the completion of the study, the audio data of each participant was processed using the PRAAT [Boersma and Weenink, 2013] toolbox to extract speaking duration, speech rate, loudness and pitch information. Overall, each participant spoke on average for 112.80 seconds during the control condition (CC) and 105.98 during the experimental condition (EC). However, this difference was not statistically significant. Nevertheless, when looking at the standard deviation (SD) of speaking duration within each group, we found significant differences between the conditions. More precisely, a paired sample T-Test revealed that the SD is significantly ($p = 0.017$) smaller in EC ($M = 30.92$) than in CC ($M = 45.07$). Similarly,

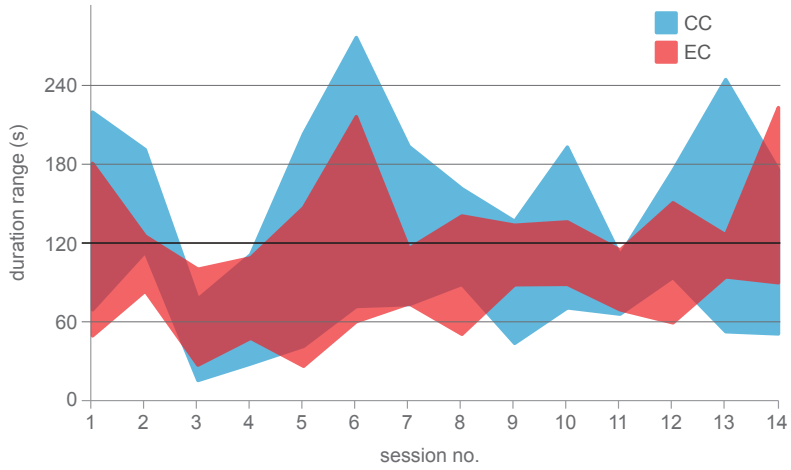

Figure 9.2: Speaking duration ranges (min - max) for all groups between conditions.

the mean maximum speaking duration over all discussion groups is significantly (p = 0.016) smaller in the EC (M = 143.53) when compared with the CC (M = 175.60). These effects can be seen in Figure 9.2 where the duration ranges (min - max) for all discussion groups are plotted. One-Way ANOVAs revealed no significant differences between the devices for both conditions in terms of speaking duration, speech rate, loudness and pitch.

The click data of the participants was also recorded and statistically analysed. When comparing between CC and EC, a paired sample T-Test revealed a significant (p = .014) increase in clicks per discussion (CC: M = .34, Sum = 24, EC: M = .97, Sum = 68). However, ANOVA tests revealed no significant differences between feedback modalities for both discussions.

Subjective Measurements

Statistical analysis on the questionnaire data also revealed interesting effects. When comparing the participants' rating against the scales midpoint (i.e. four), one-sided T-Test with applied Bonferroni-Holm correction yielded 8 significant differences. In particular, for the first research question where we investigated the effectiveness of the augmentation, one item ($M_{1\text{-}2} = 4.72, SD_{1\text{-}2} = 1.73$) was rated significantly above the middle and two items ($M_{1\text{-}1} = 3.22, SD_{1\text{-}1} = 1.63$, $M_{1\text{-}3} = 2.89, SD_{1\text{-}3} = 1.25$) were rated as significantly below the middle. No items which were meant to measure the disruption effect on the user (Q2) received ratings significantly different from the middle value. On the other hand, all items of Q3, where we measured the disruption effect on the other users, scored values significantly below the middle ($M_{3\text{-}1} = 3.02, SD_{3\text{-}1} = 1.89$, $M_{3\text{-}2} = 3.39, SD_{3\text{-}2} = 1.93$, $M_{3\text{-}3} = 2.07, SD_{3\text{-}3} = 1.31$, $M_{3\text{-}4} = 1.46, SD_{3\text{-}4} = 1.04$, $M_{3\text{-}5} = 1.91, SD_{3\text{-}5} = 1.48$).

A multivariate ANOVA revealed significant differences between the devices for Pillai's Trace, Wilks' Lambda, Hotelling's Trace and Roy's Largest Root. We followed this with univariate tests for each dependent variable. These yielded significant differences for the items i3-1 ($M_{tactile} = 2.46$, $M_{audio} = 2.15$, $M_{hmd} = 4.21$, $M_{tablet} = 3.14$) and i3-2 ($M_{tactile} = 4.00$, $M_{audio} = 2.15$, $M_{hmd} = 4.29$, $M_{tablet} = 3.07$), with participants rating the HMD as more disruptive. Post-hoc T-tests revealed significant differences only between audio and visual

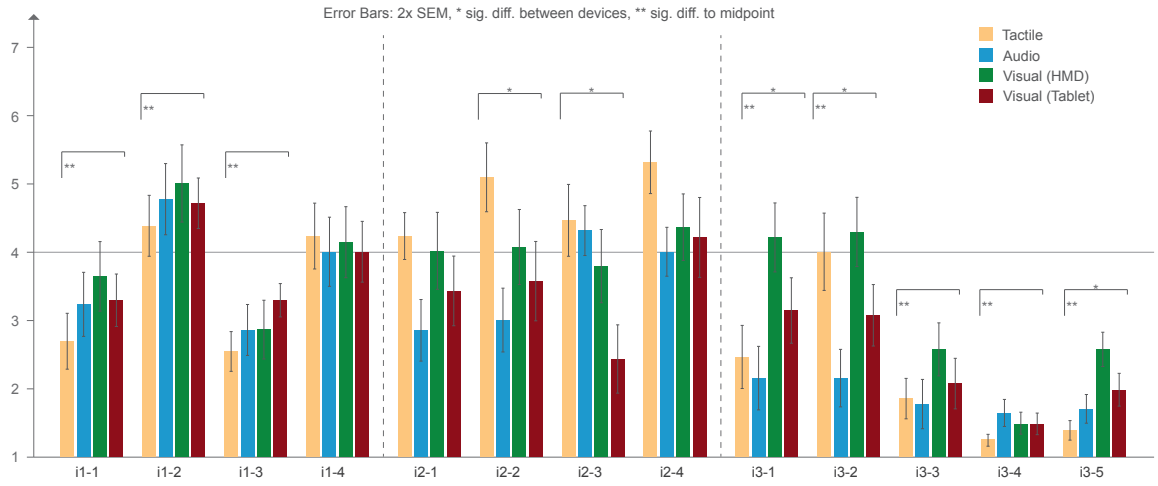

Figure 9.3: Mean values for post-session questionnaire.

HMD feedback. Significant differences between devices were also found for the items i2-2 ($M_{tactile} = 5.15$, $M_{audio} = 3.00$, $M_{hmd} = 4.07$, $M_{tablet} = 3.57$) and i2-3 ($M_{tactile} = 4.46$, $M_{audio} = 4.31$, $M_{hmd} = 3.79$, $M_{tablet} = 2.43$), with participants generally rating the tactile feedback as more disturbing. Follow-up T-tests showed significant differences only for the comparisons tactile-audio (i2-2), as well as tactile-tablet and audio-tablet (i2-3).

Furthermore, when looking at how participants rated the devices of the other interlocutors, a multivariate ANOVA revealed significant differences between the devices. Follow-up univariate tests showed significant differences for the item "The mere presence of the other person's {DEVICE} disturbed me" (i3-5) when comparing between devices (p = .001). Participants felt that the HMD worn by the other interlocutors was the most disturbing ($M_{tactile} = 1.39$, $M_{audio} = 1.70$, $M_{hmd} = 2.57$, $M_{tablet} = 1.98$). Using post-hoc T-Tests we found significant differences between HMD and tactile as well as HMD and audio. The mean values of each questionnaire item are displayed in Figure 9.3.

9.2.5 Discussion

Overall, the study yielded some valuable insights into how the social augmentation impacts the interactions. The main findings are summarized in Table 9.3.

Q1: Does the augmentation improve the social interaction? How do the feedback methods compare?

To measure the first research question, we first investigated whether the behaviour of the participants in the experimental condition (system was on) was different from that in the control condition (system was off). This showed that the behavioural feedback loops did have a positive effect on the behaviour of the users during the group discussions. More precisely, the group discussions in the experimental condition were more balanced in terms of speaking time distribution between users than the discussions in the control condition. This demonstrates that the system fulfils the first, third and fourth requirements of social augmentation (feedback was perceived and caused a behavioural change which positively effected the interaction – see Section 4.1) The effect can also be observed in Figure 9.2, where the speaking duration range for EC is "narrower" than for CC. Upon closer inspection, we

found that the augmentation had a larger impact on the more active persons, significantly reducing the maximum speaking duration across all sessions. Interestingly, this effect was not fully noticed by the participants themselves. Only one (i1-2) out of Q1's four items was rated significantly above the midpoint (four) and two were rated significantly below the midpoint. However, the fact that i1-2 was rated significantly above the midpoint tells us that the participants were able to perceive the feedback, supporting the claim that the system fulfils the first requirement of social augmentation.

When comparing between devices, we were not able to find any significant differences both in terms of objective and subjective measurements. This leads us to believe that all feedback delivery mechanisms, regardless of modality, were similar effective.

Q2: Does the augmentation disturb the user? How do the feedback methods compare?

For Q2 we first looked at the paralinguistic features and how they changed between conditions. If the augmentation disturbed the participants we would expect some differences in terms of speaking rate or voice modulation. However, we were unable to find any evidence of such an effect in either the objective or subjective data. Thus, it appears that the augmentation did not significantly disturb the user, thus fulfilling the second requirement of augmenting social interaction (Section 4.1).

Looking at the differences between devices, we were unable to find any significant differences in terms of paralinguistic features. However, the analysis of the questionnaire data revealed significant differences between the devices. When asked about the disruptiveness of the feedback devices (i2-3), post-hoc T-Tests showed that the tablet was considered the least disturbing, receiving significantly lower ratings than the Myo and headphones. This is not very surprising considering the ordinariness of the tablet and the exotic nature of the other devices. Moreover, both the Myo and the bone-conductance headphones have a fairly tight grip, potentially causing discomfort during wear.

Our analysis also revealed that the vibrotactile feedback was rated as the one they would have preferred to talk without the most (i2-2). This effect is more interesting. Since in a social interaction we rely most heavily on our vision and hearing, it would be logical for the tactile feedback to be the least disruptive and the visual and auditory feedback to be the most. However, considering the discussion in Section 6.1.3, the novelty of the feedback appears to have overshadowed the perceptual advantage of the modality. Thus, we expect that a prolonged exposure to the augmentation would reduce this effect. Also, of interest is the fact that audio feedback received the lowest ratings for i2-2 and can thus be considered the least disturbing. This stands in contrast to the results reported by Ofek et al. [2013], as they suggested that during vocal tasks, audio feedback is particularly disruptive. However, it is important to point out that in our scenario, audio feedback events were delivered in most cases only 4-5 times over a period of 2 minutes and consisted of a single short word. In Ofek's study, up to 100 words were delivered during a 5 minute window. Furthermore, Ofek did not use bone conduction speakers.

Q3: Does the augmentation disturb the other interlocutors? How do the feedback methods compare?

To measure how disturbing the device is for others, we looked at the clicker data and the questionnaires. While the clicker data is not directly able to tell us if the augmentation disturbs others, it does tell us if external persons noticed the feedback of the speaker. Looking at this

	General effects	Effects between feedback methods
Q1	Discussions were more balanced when augmentation was active	All feedback methods performed similarly well
Q2	Augmentation was not considered particularly disturbing or undisturbing by users	Tablet was considered the least disturbing device. Participants found tactile feedback to be most disturbing and audio least.
Q3	Presence of augmentation was observable by others (but was generally considered not disturbing)	The Glass' feedback was thought to be most likely to be spotted by others and external persons perceived the HMD as most disturbing device.

Table 9.3: Summary of user study findings.

data, we noticed an increase in overall amount of clicks in the experimental condition, which shows that the participants did in fact notice that the augmentation was active. Yet, there is doubt concerning the validity of this measure. The overall amount of recorded clicks was quite low and during some sessions, the clickers were not used at all. When confronted by this fact, most participants said that they were so engaged in the discussion, they forgot about the clicking task. To this end, although successfully used by Ofek et al. [2013], we were disappointed by this measurement method and thus find it difficult to recommend for other natural social interaction studies.

Moreover, the questionnaire data analysis yielded that the augmentation was not considered disruptive: All five items were rated significantly below the scale's midpoint. This suggests that the augmentation did not disturb the other interlocutors, and the system fulfils the fifth requirement of augmenting social interactions (Section 4.1).

Although, the augmentation was generally not considered disturbing, we found that some feedback methods were rated less disturbing, and others more. In particular, we found significant differences for how disruptive participants believed the feedback was (i3-1 and i3-2). Here, users generally thought that the audio feedback was the least noticeable by others, whereas the feedback delivered on the Google Glass was rated as most noticeable. When asking the opinion of the other users (i.e. the participants not wearing the device in question), we found that they felt significantly more disturbed by the Google Glass than the Myo or headphones (i3-5). These results are in line with the "glasshole" phenomenon, i.e. persons who use the Google Glass in public may be perceived as "creepy or rude." This term attracted so much interest in 2014 that Google was forced to release guidelines on how to behave when wearing the Google Glass.

9.3 Summary

The evaluation reported in this chapter shows that social augmentation also yields positive effects in face-to-face scenarios. More precisely, the proposed augmentation system improved the balance of the group discussions while not disrupting the overall flow of the conversations or negatively impacting the participants' performance. Thus, the proposed system was able to fulfil the first five requirements of the social augmentation concept (see Section 4.1). The sixth requirement (the augmentation respects the privacy of the user and the bystanders)

was not explicitly tested for. Yet, since the system only analysed the behaviour of the user which received the feedback and persistent data storing is not necessary, we would expect no privacy-related concerns in a real-world scenario.

The comparison between the four different feedback delivery methods showed no differences in terms of understandability (R1) or ability to elicit a behavioural change from the user (R3). However, although the system was generally not considered disturbing (R2), there were some differences between the devices. More specifically, the vibrotactile feedback was largely rated as most disturbing. This is in line with the novelty effect discussed in Section 6.1.3 according to which, exotic modalities are (at least initially) more disturbing than ordinary modalities. When investigating the system relative to R5 (feedback does not disturb the other interlocutors), we found that external persons rated the Google Glass as most disturbing, once again confirming the general dislike of the device.

IV Coda

10. Conclusion

Social skills training has evolved over the past few decades from using manual and analogue forms of knowledge transfer, to intelligent virtual simulation environments, which can automatically react and adapt to the learner. Now, thanks to the rapid advancement of mobile and wearable technologies seen over the past decade, a new shift in social skills training is ahead, one which allows the learner to continuously monitor and improve their social behaviour during actual social interactions.

In this context, this dissertation introduces the novel concept of *social augmentation*. Social augmentation makes use of unobtrusive wearable devices, and intelligent sensing and feedback techniques to help the user improve their behaviour while participating in social interactions. For example, a social augmentation system can help a public speaker better control their voice characteristics (speech rate, loudness) to maximize the impact on the audience. At the core of the concept lies the behavioural feedback loop, a rapid process which analyses the user's behaviour in realtime with the help of mobile social signal processing techniques, and delivers live feedback to the user on the quality of the behaviour and how to improve it. However, to achieve this, social augmentation systems must be take into account the fragile nature of human attention, the unwritten rules governing social interactions, as well as navigate various technological challenges related to the small form factor of wearable devices.

The primary goal of the dissertation was to provide the research community with a comprehensive conceptual and technological framework for social augmentation. The framework is meant to foster further research into this new and interesting domain by providing anyone with instruments to discuss, analyse and create social augmentation systems with minimal effort. A more detailed look at the contributions of the dissertation to the state-of-the-art will be provided in the following pages. Then, a discussion on possible future research directions will conclude the dissertation.

10.1 Contributions

The main contributions of this dissertation can be categorized by their nature into conceptual, technological and empirical. These are summarized in the following pages.

10.1.1 Conceptual Contributions

The main conceptual contribution is the introduction of a conceptual framework for augmenting social interactions. Social augmentation relies on the use of behavioural feedback loops, which analyse the user's behaviour and then feed it back to the user. Primarily, the feedback loop aims to generate awareness of one's own behaviour. However, through intelligent manipulation of the loop, the user can also be guided towards a more desirable behavioural state. As part of the concept definition, Chapter 4 introduced six requirements which a social augmentation system must satisfy in order to be viable and effective. The definition of the requirements has been informed by theories from the domain of cognitive psychology relating to humans' ability to distribute attention across multiple competing tasks.

For the conceptual framework, the design spaces of both behaviour analysis and feedback generation have been thoroughly explored. In terms of behaviour analysis, the dissertation promotes the use of mobile social signal processing techniques to facilitate the augmentation of live social interactions. These rely on modern wearable devices for sensing, processing and classifying user behaviour "in the wild." At the other side of the behavioural feedback loop, the use of intelligent multimodal feedback strategies is proposed. These take into consideration cognitive bottlenecks which occur in multitasking situations, and are able to automatically adapt to the user, context and scenario. Using extensive literature surveys spanning cognitive psychology, human-computer interaction and social signal processing, a total of seven types of behaviour analysis, six feedback modalities, four feedback delivery characteristics and three automatic adaptation mechanisms have been presented and measured relative to their fit in a social augmentation scenario.

The proposed conceptual framework for social augmentation gives researchers and application developers to tools necessary for making founded design choices, and facilitates the comparison between different augmentation approaches. The ability of the framework to be used as an analysis instrument has been demonstrated in Chapters 5 and 6, where two social augmentation applications have been analysed for augmentation effectiveness and obtrusiveness.

10.1.2 Technical Contributions

To support future research in the domain of social augmentation, this dissertation introduced the SSJ framework for designing and creating social augmentation systems. SSJ supports the behavioural feedback loop introduced in Section 4.2 in its entirety. On one hand, SSJ enables the recording, processing and classification of social signals in realtime on mobile devices. For this, it is able to interface with various device internal as well as external (Bluetooth-connected) sensors. On the other hand, SSJ is also capable of delivering live multimodal feedback to the user with the help of various output devices such as head-mounted displays, headphones or smart armbands. A full list of all supported sensing and output devices is provided in Appendices B and C.

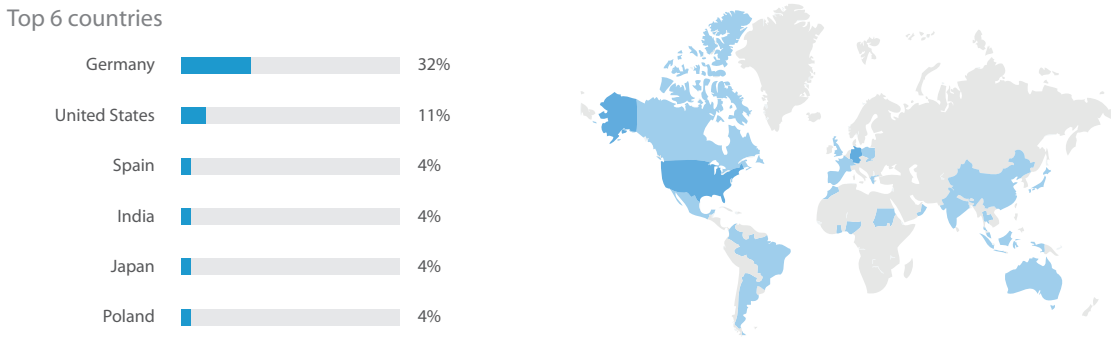

Figure 10.1: SSJ Creator installs by country between October 2016 and April 2017.

SSJ is open source and freely available for download[1]. The first version of SSJ has been released to the public in March 2016. Since then, 11 new releases spanning over 400 commits followed. SSJ is packaged as a single Android *.aar* library that can be easily integrated in Android applications. The library can either be downloaded from the homepage or added as a remote dependency in the gradle file of the application. In the latter case, the library will be automatically downloaded by the build process from the *maven* or *jcenter* repositories. Once added as a dependency to an application, the public API gives every Android application access to configuring and executing behavioural feedback loops.

Besides social augmentation, SSJ has also been successfully used as a dedicated solution for recording and analysing user behaviour in various projects and systems [Bottari, 2017; Gaibler, 2017]. One such system is Glassistant. It aims at assisting elderly with mild cognitive impairment better deal with day-to-day challenges. The application is being developed as part of a German-funded research project bearing the same name. In Glassistant, the role of SSJ is to detect when the user is in need of assistance. For example, if the user is searching for something, Glassistant uses SSJ to automatically detect the search activity, allowing it to deploy visual cues and assist the user with the search. Another use for SSJ is to detect when the user is stressed so that the Glassistant application can help them relax or notify a family member.

To improve the accessibility of SSJ, the dissertation also introduced the *SSJ Creator* Android application. It allows persons without a technical background to work with SSJ using a modern graphical user interface. More precisely, it enables the design and execution of SSJ pipelines without writing a single line of code. The application has been added to the Google Play Store in October 2016. Since then, it has been downloaded more than 80 times by users from over 10 countries (see Figure 10.1), accounting to an average of roughly 10 downloads per month. Out of the 80 downloads, over 30 installations are still active. This high conversion rate and the broadness of the install base suggests that the application is useful for the general public as well, despite its roots as a research instrument.

10.1.3 Empirical Contributions

The social augmentation concept has been empirically tested with the help of three user studies. The first two studies measured the potential of a social augmentation system for improving public speaking related behaviour (see Chapter 8). While both studies made use

[1] `http://hcm-lab.de/ssj/`

of realistic setups, asking the users to use the social augmentation systems while holding a presentation, the second study had participants use the system during their actual presentations at an annual workshop. Overall, the studies found that the social augmentation system significantly improved the quality of the participants' behaviour, and that both the users and the interlocutors rapidly grew accustomed to the system, despite a very bulky setup. Moreover, users were generally positive regarding the usefulness of the system, with many expressing interest in using such a system outside of the experiment.

For the third study, we investigated the effectiveness and disruptiveness of four different types of social augmentation in a face-to-face group discussion scenario (see Chapter 9). More specifically, with the help of 54 participants, we compared between visual, auditory and tactile feedback for helping the users maintain a more balanced group discussion. The study yielded that, regardless of feedback method, the augmentation helped improve the balance of the group discussion while not disturbing the users or the interlocutors. Still, there were differences between feedback methods regarding their disruptiveness, with users generally favouring ordinary feedback devices (tablet, headphones) over the more exotic ones (vibrotactile armband, Google Glass). This suggests that designers of augmentation systems should attempt to choose conventional feedback mechanics over more exotic ones, or take an accommodation period in consideration to reduce the disruptive effects of unfamiliar feedback (see discussion in Section 6.1).

Besides these primary evaluations, smaller studies have also been carried out and were aimed at informing the development of the social augmentation concept. These studies can be seen as evolutionary precursors and helped shape the final social augmentation concept presented in this thesis. In this regard, the TARDIS study introduced in Section 3.1 had the largest impact on the development of social augmentation concept. It demonstrated the power of automated behaviour analysis and live feedback for use in a training environment. Although the TARDIS system was stationary and relied on an imperfect virtual simulation as its main training instrument, it showed that social skill training can be made more accessible and scalable with the help of off-the-shelf technology. Thus, it motivated the development of the social augmentation concept.

Overall, the empirical findings have not only proven the effectiveness of the proposed concept, but have also provided evidence that technology is feasible to be deployed in a social context if correctly designed. More specifically, systems which aim to work alongside social interactions need to be mindful of the fragile nature of human attention and respectful of the rules governing social interactions, such as user and bystander privacy.

10.2 Future Work

Social augmentation as a type of social skills training method is still in its incipient phase and further research is needed to better explore its abilities.

10.2.1 External Signals

In this dissertation we explicitly constrained social augmentation to analysing only the behaviour of the user. In a world where our right to privacy is threatened by an ever more invasive use of technology, this design choice acts as a layer of protection meant to prevent misuse of the social augmentation concept. Yet, when participating in social interactions, our behaviour is seldom the outcome of only internal processes as we continuously react and

adapt to the actions of others. Thus, allowing the augmentation to draw information from both the user and the interlocutors would most likely benefit the quality of the augmentation. For example, while performing a public speech, the system could analyse both the user's gestures and the audience's reaction to the gestures, to deliver more appropriate feedback. The augmentation could also take into consideration interpersonal social processes such as mimicry [Chartrand and Bargh, 1999] or emotion regulation [Kappas, 2013]. Consider two dyadic conversations between peers which are identical but for the emotional state of the interlocutor. In one, the interlocutor is calm and speaks softly, whereas in the other, the interlocutor is more energetic and speaks loudly. Providing the user with the same behavioural feedback (e.g. speak louder) in both situations will likely lead to unsatisfactory results. An augmentation system which is also aware of the interlocutors' behaviour could more accurately adapt the feedback to each situation.

Externally-fed feedback loops could also be used to help persons with disabilities better react to their interlocutors. For example, persons on the autism spectrum, which are known to have issues with interpreting social signals [Bolte and Poustka, 2003; Braverman et al., 1989; Capps et al., 1992; Sigman et al., 2006], could be provided with information on the meaning of the facial expressions, gestures or postures of others. This could improve their ability to integrate in social groups or navigate social situations. One example of such a system is the facial expression sonification developed by my colleagues and I [Dietz et al., 2016]. Here, we used the front-facing camera of a mobile eye-tracker to provide visually impaired users with information on the facial expressions of their interlocutors.

However, using other persons as signal sources for the behavioural analysis does raise some very serious privacy concerns. Do these other persons need to know that their signals are used? Would knowing it change how they behave? Do they need to consent to it? If yes, would they do it?

10.2.2 Long-Term Studies

The user studies introduced in this dissertation have focused mainly on the immediate effects of behavioural feedback loops. However, the question arises of what longer-term effects social augmentation systems have on the user. In this regard, two directions are of most interest: automation of feedback reaction and long-term learning.

First, as discussed in Section 2.2.4, it is possible to automate the execution of certain tasks through repetition. Thus, one matter worth exploring is the degree to which the reaction to a particular feedback can be automated through repeated exposure. For instance, a person which is prone to speaking too fast could be provided with an augmentation system which provides subtle vibrotactile feedback every time behaviour analysis detects an excessive speaking rate. Now, the interesting question is: If the person uses the augmentation system over a longer period, does the reaction to the feedback (speaking slower) become automatic, i.e. is it triggered subconsciously, or does the user still need to consciously decide to reduce the speaking rate.

The second effect long-term studies could help investigate is the actual learning of social behaviour. More specifically, such user studies could measure whether social augmentation causes a long-term improvement of user social skills so that, after a longer exposure, the augmentation is not needed any more. Finding evidence of such an effect would further strengthen the claim of the social augmentation concept as an alternative to classical social skills training approaches.

10.2.3 Social Interaction Classification

Section 4.3.1 discussed the idea of having an augmentation system automatically activate as soon as the user engages in a social interactions. However, current approaches on social interaction classification are limited to simply detecting whether a social interaction is happening or not. For social augmentation however, it would be of benefit to not only classify whether a social interaction is happening, but also the type of social interaction. This would allow the augmentation system to automatically load the feedback strategy appropriate to the current type of interaction.

To achieve this, additional sensors could be employed to also record background noise, the user's speech or even point of view. Baur et al. [2015] propose using eye gaze data for differentiating between different types of gaze. Such data could enable a more detailed classification of the social interaction, e.g. by differentiating between public speaking, social conversations and business conversations. However, as previously discussed, it is important to note that analysing external audio and video signals could raise privacy issues. Thus, the fit of such techniques for social augmentation must be investigated closely.

10.2.4 Timing Management

Current timing management techniques used in social augmentation systems are quite basic and mostly involve reducing the frequency of possible interruptions to the primary task. Yet, as discussed in Section 4.3.3, there is ongoing research on automatically inferring the cost of an interruption. Such approaches could be used in social augmentation systems to more actively choose the most opportune moment at which feedback can be delivered. For instance, if the behaviour analysis notices the user is speaking too fast but he is in the middle of an explanation, feedback could be delayed until after the user finishes the explanation or gives the turn to the interlocutor. Such a technique would not only reduce the probability of interrupting the primary task, but also make the feedback events more likely to be correctly perceived and to trigger a reaction.

The problem which arises is that unlike the scenarios used by Horvitz and Apacible [2003] (office activity) or Poppinga et al. [2014] (phone activity throughout the day), opportune moments for social augmentation are much shorter, ranging from several hundred milliseconds (pauses between sentences) to several seconds (switching slides, passing the turn). Accurately identifying these, is a much more challenging undertaking. One approach would be to combine the approach used by Horvitz and Poppinga with more accurate sensing hardware, such as eye trackers. Eye trackers allow a more detailed analysis of the user's focus of attention, and thus might help better classify the shorter opportune moments found in a social augmentation context. Another approach would be to use audio processing for detecting pauses in continuous speech. Based on the paralinguistical characteristics of the preceding utterance, as well as the length of the pause, the system could decide whether the pause is in fact appropriate for feedback delivery.

10.2.5 Mobile Machine Learning

Since the aim of social augmentation is to be used "in the wild" and during actual interactions, one possible improvement would be to allow the behaviour analysis component to continuously update its classification models from data collected while in use. Such a technique would enable the behaviour analysis to adapt to the user, and thus learn to better detect inappropriate behaviour. For instance, a social augmentation system tasked with helping the

user manage stress could occasionally ask them to annotate whether they are feeling stressed or not. The behaviour analysis component would then store this information and use it to improve its stress classification model with the help of online learning mechanisms. This would allow it to tailor the classification to the needs and physiology of the current user, thus increasing detection accuracy.

Online learning would be a perfect fit for SSJ as it would only strengthen its ability to classify human behaviour in a mobile context. One possible approach is to still use offline trained models, but allow them to be continuously extended and updated using new data. For this, SSJ Creator's live annotation features could be used to enable the user to label the data on-the-go using their smart armband (see Section 7.6.3).

Active learning mechanisms could also be employed to ask the user for an annotation every time "interesting" data is detected. For example, in the case of the *query by uncertainty* pattern [Lewis and Gale, 1994; Settles, 2012], when the confidence level of an active classification task drops below a certain threshold, the system could ask the user to annotate the data which caused this classification result.

For the annotation requests, the feedback delivery mechanics discussed in Chapters 4 and 6 could be used to minimize the likelihood of disrupting the user's primary task.

10.2.6 Additional Application Scenarios

Finally, social augmentation is by no means limited to public speaking or group discussions. One possibility would be to use social augmentation for improving user behaviour during job interviews. Job interviews, as a type of human-human interaction, rely on the ability of the interlocutors to read each other's behaviour and emotion. For the recruiter, the goal is to determine whether the interviewee is adequate for a specific job [Posthuma et al., 2002]. In this scenario, it is expected from the interviewee to remain composed and not show extreme emotions [Sieverding, 2009]. More precisely, behaviours such as body expressivity (e.g. gestures, postures or facial expressions), vocal quality (e.g. speech rate, loudness, etc.) and eye gaze behaviour have been found to play a significant role during job interviews [Hollandsworth et al., 1979; Carl, 1980]. Considering this, the interviewee's nonverbal behaviour is a crucial element in the outcome of the whole interview and an increased awareness would only benefit this outcome. Thus, social augmentation would be a good fit to help the user better regulate their behaviour. Due to the delicate nature of such an interaction, the choice in sensing and output devices is critical. For example, a smart watch or smart armband allows the inconspicuous analysis of the user's gesture while also permitting the delivery of subtle, low intensity vibrotactile feedback.

Another possible scenario represents information-sensitive conversations, such as physician-patient conversations, where ensuring sufficient behavioural awareness can be advantageous to the interaction goals. Blanch et al. [2009] showed that the quality of nonverbal skills impacts how the physician is perceived by the patient. Here, social augmentation can help the physician with the delivery of sensitive information. For example, the system could remind the physician to use pauses to allow the patient to process the information, or provide feedback to increase vocal quality (e.g. speech rate, loudness) and boost the likelihood of the patient correctly understanding the message.

Mixed-culture interactions are another use case for social augmentations. Cultural misunderstandings are common and can significantly impact the outcome of a social interaction [Endrass, 2014; Ting-Toomey, 1999]. Among others, the appropriate use of gestures, personal

space and turn taking behaviour are known to vary between cultures. An augmentation system deployed in such scenarios could help businessmen during negotiations in foreign countries, but also tourists who wish to avoid misunderstandings.

Lastly, social augmentation may also be used in a therapeutic context to help persons with disabilities better navigate social interactions. Some examples have already been provided for assisting persons on the autism spectrum [Boyd et al., 2016] or those suffering from Parkinson's disease [McNaney et al., 2015], but there is still room for further research. Applying social augmentation is such scenarios would help empower the less fortunate and contribute to undermining societal divisions.

Bibliography

Abramov, I., Gordon, J., and Chan, H. (1991). Color appearance in the peripheral retina: Effects of stimulus size. *Journal of the Optical Society of America A*, 8(2):404.

Abravanel, E. (1971). Active detection of solid-shape information by touch and vision. *Perception & Psychophysics*, 10(5):358–360.

Adib, F., Mao, H., Kabelac, Z., Katabi, D., and Miller, R. C. (2015). Smart homes that monitor breathing and heart rate. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 837–846. ACM.

Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659.

Alon, J., Athitsos, V., Yuan, Q., and Sclaroff, S. (2009). A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1685–1699.

Altmann, E. M. and Trafton, J. G. (2004). Task interruption: Resumption lag and the role of cues. In *Cognitive Science Society, Conference Proceedings*, pages 43–48.

Altmann, E. M., Trafton, J. G., and Hambrick, D. Z. (2014). Momentary interruptions can derail the train of thought. *Journal of Experimental Psychology: General*, 143(1):215.

Amoore, J. E., Johnston, J. W., and Rubin, M. (1964). The stereochemical theory of odor. *Scientific American*, 210(2):42–49.

Anderson, B. A., Laurent, P. A., and Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences (PNAS)*, 108(25):10367–10371.

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., and Sabouret, N. (2013). The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment (ACE), Conference Proceedings*, volume 8253 of *Lecture Notes in Computer Science*, pages 476–491. Springer.

Antoniou, A., Theodoridis, E., Chatzigiannakis, I., and Mylonas, G. (2011). Monitoring physical space using mobile phones for inferring social and contextual interactions. In *Sensors, Conference Proceedings*, pages 1616–1619. IEEE.

Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze*. Cambridge Univ. Press, Cambridge.

Armstrong, N. and Wagner, M. (2003). *Field guide to gestures: How to identify and interpret virtually every gesture known to man*. Quirk, Philadelphia, Pa.

Arroyo, E. and Selker, T. (2003). Self-adaptive multimodal-interruption interfaces. In *Intelligent User Interfaces, Conference Proceedings*, pages 6–11. ACM.

Arroyo, E., Selker, T., and Stouffs, A. (2002). Interruptions as multimodal outputs: Which are the less disruptive? In *Multimodal Interfaces (ICMI), Conference Proceedings*. IEEE.

Asif, A. and Boll, S. (2010). Where to turn my car?: comparison of a tactile display and a conventional car navigation system under high load condition. In *Automotive User Interfaces and Interactive Vehicular Applications, Conference Proceedings*. ACM.

Awadeen, M. (2015). *Investigating Potential Distraction of Head Mounted Displays*. Bachelor's thesis, supervised by Damian, I., and André, E., Universität Augsburg, Augsburg, Germany.

Aylett, R., Hall, L., Tazzyman, S., Endrass, B., André, E., Ritter, C., Nazir, A., Paiva, A., Höfstede, G., and Kappas, A. (2014). Werewolves, cheats, and cultural sensitivity. In *Autonomous agents and multi-agent systems (AAMAS), Conference Proceedings*, pages 1085–1092. IFAAMAS.

Bailenson, J. N. and Yee, N. (2005). Digital chameleons: automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819.

Bailey, B. P. and Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(4):21.

Baker, L. E. (1937). The influence of subliminal stimuli upon verbal behavior. *Journal of Experimental Psychology*, 20(1):84.

Baker, L. E. (1938). The pupillary response conditioned to subliminal auditory stimuli. *Psychological Monographs*, 50(3):i.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Banf, M. and Blanz, V. (2013). Sonification of images for the visually impaired using a multi-level approach. In *Augmented Human (AH), Conference Proceedings*. ACM.

Bannach, D., Amft, O., and Lukowicz, P. (2008). Rapid prototyping of activity recognition applications. *Pervasive Computing*, 7(2):22–31.

Barmaki, R. and Hughes, C. E. (2015). Providing real-time feedback for student teachers in a virtual rehearsal environment. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Barnlund, D. C. (1975). *Public and private self in Japan and the United States: Communicative styles of two cultures*. Simul Press.

Barral, O., Aranyi, G., Kouider, S., Lindsay, A., Prins, H., Ahmed, I., Jacucci, G., Negri, P., Gamberini, L., Pizzi, D., and Cavazza, M. (2014). Covert persuasive technologies: Bringing subliminal cues to human-computer interaction. In *Persuasive Technology, Conference Proceedings*, volume 8462 of *Lecture Notes in Computer Science*, pages 1–12. Springer, Cham.

Bartoshuk, L. M. (2000). Comparing sensory experiences across individuals: Recent psychophysical advances illuminate genetic variation in taste perception. *Chemical Senses*, 25(4):447–460.

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., and Scherer, S. (2013). Cicero - towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, volume 8108 of *Lecture Notes in Computer Science*, pages 116–128. Springer Berlin Heidelberg, Berlin, Heidelberg.

Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., and André, E. (2013a). A job interview simulation: Social cue-based interaction with a virtual character. In *Social Computing, Conference Proceedings*, pages 220–227. IEEE.

Baur, T., Damian, I., Lingenfelser, F., Wagner, J., and André, E. (2013b). Nova: Automated analysis of nonverbal signals in social interactions. In *Human Behavior Understanding, Workshop Proceedings*, volume 8212 of *Lecture Notes in Computer Science*. Springer.

Baur, T., Mehlmann, G., Damian, I., Lingenfelser, F., Wagner, J., Lugrin, B., André, E., and Gebhard, P. (2015). Context-aware automated analysis and annotation of social human–agent interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(2):11.

Bayle, D. J., Henaff, M.-A., and Krolak-Salmon, P. (2009). Unconsciously perceived fear in peripheral vision alerts the limbic system: a MEG study. *PloS one*, 4(12):e8207.

Benbasat, I. and Todd, P. (1993). An experimental investigation of interface design alternatives: Icon vs. text and direct manipulation vs. menus. *International Journal of Man-Machine Studies*, 38(3):369–402.

Bendini, S. A. (1964). Holy smoke: The oriental fireclocks. *New Scientist*, (21):537–539.

Berglund, B., Berglund, U., Engen, T., and Lindvall, T. (1971). The effect of adaptation on odor detection. *Perception & Psychophysics*, 9(5):435–438.

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. McGraw-Hill Book Company, New York, NY, US.

Bernsen, N. O. (1994). Modality theory in support of multimodal interface design. In *Intelligent Multi-Media Multi-Modal Systems, Symposium Proceedings*, pages 37–44. AAAI.

Bernsen, N. O. (2008). Multimodality theory. In *Multimodal User Interfaces*, Signals and Commmunication Technologies, pages 5–29. Springer, Berlin, Heidelberg.

Bertin, J. (1983). *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press.

Birdwhistell, R. L. (2011). *Kinesics and context: Essays on body motion communication*. University of Pennsylvania press.

Blanch, D. C., Hall, J. A., Roter, D. L., and Frankel, R. M. (2009). Is it good to express uncertainty to a patient? correlates and consequences for medical students in a standardized patient visit. In *Communication in Healthcare (EACH), Conference Proceedings*, volume 76 of *Patient Education and Counseling*, pages 300–306. Elsevier.

Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M. (1989). Earcons and icons: their structure and common design principles. *Human-Computer Interaction*, 4(1):11–44.

Bliss, J. C., Crane, H. D., Mansfield, P. K., and Townsend, J. T. (1966). Information available in brief tactile presentations. *Perception & Psychophysics*, 1(4):273–283.

Bodnar, A., Corbett, R., and Nekrasovski, D. (2004). Aroma: ambient awareness through olfaction in a messaging application. In *Multimodal Interfaces (ICMI), Conference Proceedings*, pages 183–190. ACM.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer. http://www.praat.org/.

Bologna, G., Deville, B., and Pun, T. (2009). On the use of the auditory pathway to represent image scenes in real-time. *Neurocomputing*, 72(4-6):839–849.

Bolte, S. and Poustka, F. (2003). The recognition of facial affect in autistic and schizophrenic subjects and their first-degree relatives. *Psychological medicine*, 33(5):907–915.

Bornstein, R. F. (1990). Critical importance of stimulus unawareness for the production of subliminal psychodynamic activation effects: A meta–analytic review. *Journal of Clinical Psychology*, 46(2):201–210.

Bottari, M. (2017). *Subliminale Perzeption von computergeneriertem Feedback in einem sozialen Kontext*. Master thesis, supervised by Damian, I., and André, E., Universität Augsburg, Augsburg, Germany.

Boyd, L. E., Rangel, A., Tomimbang, H., Conejo-Toledo, A., Patel, K., Tentori, M., and Hayes, G. R. (2016). Saywat: Augmenting face-to-face conversations for adults with autism. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 4872–4883. ACM.

Brainard, M. S. and Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature reviews. Neuroscience*, 1(1):31–40.

Braverman, M., Fein, D., Lucci, D., and Waterhouse, L. (1989). Affect comprehension in children with pervasive developmental disorders. *Journal of autism and developmental disorders*, 19(2):301–316.

Brewster, S. and Brown, L. M. (2004). Tactons: structured tactile messages for non-visual information display. In *Australasian User Interface, Conference Proceedings*. Australian Computer Society.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, 64(3):205–215.

Bronkhorst, A. W., Veltman, J. A., and van Breda, L. (1996). Application of a three-dimensional auditory display in a flight task. *Human Factors*, 38(1):23–33.

Brooks, L. R. (1968). Spatial and verbal components of the act of recall. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 22(5):349–368.

Brooks, S. J., Savov, V., Allzén, E., Benedict, C., Fredriksson, R., and Schiöth, H. B. (2012). Exposure to subliminal arousing stimuli induces robust activation in the amygdala, hippocampus, anterior cingulate, insular cortex and primary visual cortex: a systematic meta-analysis of fmri studies. *NeuroImage*, 59(3):2962–2973.

Bulling, A., Roggen, D., and Tröster, G. (2009). Wearable eog goggles: Seamless sensing and context-awareness in everyday environments. *Journal of Ambient Intelligence and Smart Environments*, 1(2):157–171.

Burke, J. L., Prewett, M. S., Gray, A. A., Yang, L., Stilson, F. R. B., Coovert, M. D., Elliot, L. R., and Redden, E. (2006). Comparing the effects of visual-auditory and visual-tactile feedback on user performance: a meta-analysis. In *Multimodal Interfaces (ICMI), Conference Proceedings*. ACM.

Cades, D. M., Trafton, J. G., and Boehm-Davis, D. A. (2006). Mitigating disruptions: Can resuming an interrupted task be trained? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(3):368–371.

Campbell, A., Choudhury, T., Hu, S., Lu, H., Mukerjee, M. K., Rabbi, M., and Raizada, R. D. (2010). Neurophone: brain-mobile phone interface using a wireless eeg headset. In *Workshop on Networking, systems, and applications on mobile handhelds*.

Campbell, R., Dodd, B., and Burnham, D. K. (1998). *Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual speech*. Psychology Press, Hove, East Sussex, UK.

Capps, L., Yirmiya, N., and Sigman, M. (1992). Understanding of simple and complex emotions in non-retarded children with autism. *Journal of child psychology and psychiatry, and allied disciplines*, 33(7):1169–1182.

Carbonaro, N., Anania, G., Mura, G. D., Tesconi, M., Tognetti, A., Zupone, G., and de Rossi, D. (2011). Wearable biomonitoring system for stress management: A preliminary study on robust ecg signal processing. In *A World of Wireless, Mobile and Multimedia Networks, Conference Proceedings*, pages 1–6. IEEE.

Caridakis, G., Raouzaiou, A., Karpouzis, K., and Kollias, S. (2006). Synthesizing gesture expressivity based on real sequences. *Journal of Multimodal Corpora*.

Carl, H. (1980). Nonverbal communication during the employment interview. *Business Communication Quarterly*, 43(4):14–19.

Carlson, D. and Schrader, A. (2012). Dynamix: An open plug-and-play context framework for android. In *Internet of Things (IOT), Conference Proceedings*, pages 151–158. IEEE.

Carreras, I., Matic, A., Saar, P., and Osmani, V. (2012). Comm2sense: Detecting proximity through smartphones. In *Pervasive Computing and Communications Workshops, Conference Proceedings*. IEEE.

Chang, A., Resner, B., Koerner, B., Wang, X., and Ishii, H. (2001). Lumitouch: an emotional communication device. In *Human Factors in Computing Systems (CHI), Conference Proceedings*. ACM.

Chang, K.-h., Fisher, D., Canny, J., and Hartmann, B. (2011). How's my mood and stress?: an efficient speech analysis library for unobtrusive monitoring on mobile phones. In *Body Area Networks, Conference Proceedings*, pages 71–77. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Chen, J. and Engelen, L. (2012). *Food Oral Processing*. Wiley-Blackwell, Oxford, UK.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975.

Cholewiak, R. W. (1979). Spatial factors in the perceived intensity of vibrotactile patterns. *Sensory Processes*.

Chollet, M., Stefanov, K., Prendinger, H., and Scherer, S. (2015). Public speaking training with a multimodal interactive virtual audience framework. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Ciolek, T. M. and Kendon, A. (1980). Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry*, 50(3-4):237–271.

Claus, D., Hilz, M. J., Hummer, I., and Neundorfer, B. (1987). Methods of measurement of thermal thresholds. *Acta neurologica Scandinavica*, 76(4):288–296.

Cohen, B. M. and Etheredge, J. M. (1975). Recruiting's main ingredient. *Journal of College Placement*.

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., and Young, R. M. (1995). Four easy pieces for assessing the usability of multimodal interaction: The care properties. In *Human-Computer Interaction (INTERACT), Conference Proceedings*, Advances in Information and Communication Technology, pages 115–120. Springer.

Cristian, F. (1989). Probabilistic clock synchronization. *Distributed Computing*, 3(3):146–158.

Curcio, C. A., Sloan, K. R., Kalina, R. E., and Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, 292.

Cutrell, E., Czerwinski, M., and Horvitz, E. (2001). Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Human-Computer Interaction (INTERACT), Conference Proceedings*. IOS Press.

Cutrell, E. B., Czerwinski, M., and Horvitz, E. (2000). Effects of instant messaging interruptions on computing tasks. In *Human Factors In Computing Systems (CHI), Conference Proceedings*. ACM.

Czerwinski, M., Cutrell, E., and Horvitz, E. (2000). Instant messaging: Effects of relevance and timing. In *People and computers XIV: Proceedings of HCI*, volume 2. British Computer Society.

Damian, I. (2012). *Customizing Agent Interactions*. Diploma thesis, Augsburg, Germany.

Damian, I. and André, E. (2016). Exploring the potential of realtime haptic feedback during social interactions. In *Tangible, Embedded, and Embodied Interaction (TEI), Conference Proceedings*, pages 410–416. ACM.

Damian, I., Baur, T., and André, E. (2013a). Investigating social cue-based interaction in digital learning games. In *Intelligent Digital Games for Empowerment and Inclusion (IDGEI), satelite Workshop to Foundations of Digital Games (FDG), Workshop Proceedings*. SASDG.

Damian, I., Baur, T., and André, E. (2016a). Measuring the impact of behavioural feedback loops on social interactions. In *Multimodal Interaction (ICMI), Conference Proceedings*, pages 201–208. ACM.

Damian, I., Baur, T., Lugrin, B., Gebhard, P., Mehlmann, G., and André, E. (2015a). Games are better than books: In-situ comparison of an interactive job interview game with conventional training. In *Artificial Intelligence in Education (AIED), Conference Proceedings*, volume 9112 of *Lecture Notes in Computer Science*, pages 84–94. Springer, Cham.

Damian, I., Baur, T., Tan, C. S. S., Schöning, J., Luyten, K., and André, E. (2014a). Towards peripheral feedback-based realtime social behaviour coaching. In *Interactions and Applications on See-through Technologies, Workshop Proceedings*.

Damian, I., Dietz, M., Gaibler, F., and André, E. (2016b). Social signal processing for dummies. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Damian, I., Endrass, B., Bee, N., and André, E. (2011a). A software framework for individualized agent behavior. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, volume 6895 of *Lecture Notes in Computer Science*, pages 437–438. Springer.

Damian, I., Endrass, B., Huber, P., Bee, N., and André, E. (2011b). Individualizing agent interactions. In *Motion in Games (MIG), Conference Proceedings*. Springer.

Damian, I., Janowski, K., and Sollfrank, D. (2009). Spectators, a joy to watch. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, volume 5773 of *Lecture Notes in Computer Science*, pages 558–559. Springer.

Damian, I., Kistler, F., Obaid, M., Buhling, R., Billinghurst, M., and Andre, E. (2013b). Motion capturing empowered interaction with a virtual agent in an augmented reality environment. In *Mixed and Augmented Reality (ISMAR), Conference Proceedings*, pages 1–6. IEEE.

Damian, I., Obaid, M., Kistler, F., and André, E. (2013c). Augmented reality using a 3d motion capturing suit. In *Augmented Human (AH), Conference Proceedings*, pages 233–234. ACM.

Damian, I., Tan, C. S., Baur, T., Schöning, J., Luyten, K., and André, E. (2015b). Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 565–574. ACM.

Damian, I., Tan, C. S. S., Baur, T., Schöning, J., Luyten, K., and André, E. (2014b). Exploring social augmentation concepts for public speaking using peripheral feedback and real-time behavior analysis. In *Mixed and Augmented Reality (ISMAR), Conference Proceedings*. IEEE.

Davies, T. and Beeharee, A. (2012). The case of the missed icon: change blindness on mobile devices. In *Human Factors In Computing Systems (CHI), Conference Proceedings*. ACM.

de Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences*, 10(5):204–211.

Dermody, F. and Sutherland, A. (2015). A multimodal system for public speaking with real time feedback. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Dermody, F. and Sutherland, A. (2016). Multimodal system for public speaking with real time feedback: a positive computing perspective. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70(1):80.

DeVaul, R. W., Pentland, A., and Corey, V. R. (2003). The memory glasses: Subliminal vs. overt memory support with imperfect information. In *Wearable Computers (ISWC), Conference Proceedings*. IEEE.

Dietz, M., El Garf, M., Damian, I., and André, E. (2016). Exploring eye-tracking-driven sonification for the visually impaired. In *Augmented Human (AH), Conference Proceedings*. ACM.

Dmitrenko, D., Vi, C. T., and Obrist, M. (2016). A comparison of scent-delivery devices and their meaningful use for in-car olfactory interaction. In *Automotive User Interfaces and Interactive Vehicular Applications, Conference Proceedings*. ACM.

Dodd, B. (1977). The role of vision in the perception of speech. *Perception*, 6(1):31–40.

Drake, L. R., Kaplan, H. R., and Stone, R. A. (1972). How do employers value the interview? *Journal of College Placement*, 32(3):47–51.

Dwyer, K. K. and Davidson, M. M. (2012). Is public speaking really more feared than death? *Communication Research Reports*, 29(2):99–107.

Eagle, N. and Pentland, A. (2005). Social serendipity: Mobilizing social software. *IEEE Pervasive Computing*, 4(2):28–34.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Times Books, New York, 1st ed. edition.

Ekman, P. (2009). Lie catching and microexpressions. In *The Philosophy of Deception*.

Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica*, 1(1):49–98.

Ekman, P. and Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Prentice-Hall.

Endrass, B. (2014). *Cultural diversity for virtual characters: Investigating behavioral aspects across cultures*. Springer, Wiesbaden.

Endrass, B., Damian, I., Huber, P., Rehm, M., and Andre, E. (2010). Generating culture-specific gestures for virtual agent dialogs. In *Intelligent Virtual Agents (IVA), Conference Proceedings*, volume 6356 of *Lecture Notes in Computer Science*, pages 329–335. Springer.

Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. *Air Force Office of Scientific Research*, (3223).

Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., and Shah, S. (2011). Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Embedded Networked Sensor Systems, Conference Proceedings*, pages 274–287. ACM.

Fairbanks, G. (1955). Selective vocal effects of delayed auditory feedback. *Journal of Speech & Hearing Disorders*.

Feese, S., Arnrich, B., Troster, G., Burtscher, M., Meyer, B., and Jonas, K. (2013). Coenofire: monitoring performance indicators of firefighters in real-world missions using smartphones. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*. ACM.

Ferraro, G. (1990). *The cultural dimension of international business*. Prentice Hall.

Flutura, S., Wagner, J., Lingenfelser, F., Seiderer, A., and André, E. (2016). MobileSSI: Asynchronous fusion for social signal interpretation in the wild. In *Multimodal Interaction (ICMI), Conference Proceedings*, pages 266–273. ACM.

Freides, D. (1974). Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit. *Psychological bulletin*, 81(5):284.

Fridlund, A. J. (2014). *Human Facial Expression: An Evolutionary View*. Elsevier Science.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., and Landay, J. A. (2007). Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Mobile Systems, Applications and Services (MobiSys), Conference Proceedings*. ACM.

Fukumoto, K., Terada, T., and Tsukamoto, M. (2013). A smile/laughter recognition mechanism for smile-based life logging. In *Augmented Human (AH), Conference Proceedings*, pages 213–220. ACM.

Fung, M., Jin, Y., Zhao, R., and Hoque, M. (2015). Roc speak: semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*. ACM.

Gabbard, J. L., Swan, J. E., Zedlitz, J., and Winchester, W. W. (2010). More than meets the eye: An engineering study to empirically examine the blending of real and virtual color spaces. In *Virtual Reality (VR), Conference Proceedings*, pages 79–86. IEEE.

Gaggioli, A., Pioggia, G., Tartarisco, G., Baldus, G., Ferro, M., Cipresso, P., Serino, S., Popleteev, A., Gabrielli, S., Maimone, R., and Riva, G. (2012). A system for automatic detection of momentary stress in naturalistic settings. *Studies in health technology and informatics*, 181:182–186.

Gaibler, F. (2017). *Erkennung von emotionalem Stress mit Hilfe von mobilen Geräten*. Master thesis, supervised by Damian, I., and André, E., Universität Augsburg.

Gallaher, P. (1992). Individual differences in nonverbal behavior: Dimensions of style. *Journal of personality and social psychology*, 63.

Gaver, W. (1986). Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2(2):167–177.

Gaver, W. W. (1989). The sonicfinder: an interface that uses auditory icons. *Human-Computer Interaction*, 4(1):67–94.

Gebhard, P., Baur, T., Damian, I., Mehlmann, G., Wagner, J., and André, E. (2014). Exploring interaction strategies for virtual characters to induce stress in simulated job interviews. In *Autonomous Agents and Multi-Agent Systems (AAMAS), Conference Proceedings*, pages 661–668. IFAAMAS.

Gebhard, P., Mehlmann, G., and Kipp, M. (2012). Visual scenemaker - a tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 6(1-2):3–11.

Ghose, A., Bhaumik, C., and Chakravarty, T. (2013). Blueeye: a system for proximity detection using bluetooth on mobile phones. In *Pervasive and Ubiquitous Computing (UbiComp), Workshop Proceedings*. ACM.

Gibbons, B. (1986). The intimate sense of smell. *National Geographic*, (170):324–361.

Gillie, T. and Broadbent, D. (1989). What makes interruptions disruptive? a study of length, similarity, and complexity. *Psychological Research*, 50(4):243–250.

Gluhak, A., Presser, M., Zhu, L., Esfandiyari, S., and Kupschick, S. (2007). Towards mood based mobile services and applications. In *Smart Sensing and Context, Conference Proceedings*, pages 159–174. Springer.

Goldstein, A. P., Carr, E. G., and Davidson, W. S. (2013). *In Response to Aggression: Methods of Control and Prosocial Alternatives*. Elsevier Science.

Gooch, D. and Watts, L. (2010). Communicating social presence through thermal hugs. In *Social Interaction in Spatially Separated Environments, Workshop Proceedings*. University of Bath.

Greenwald, A. G. (1992). New look 3: Unconscious cognition reclaimed. *American Psychologist*, 47(6):766.

Guizatdinova, I. and Guo, Z. (2003). Sonification of facial expressions. *New Interaction Techniques*, page 44.

Gumperz, J. J. (1982). *Discourse strategies*, volume 1 of *Studies in interactional sociolinguistics*. Cambridge Univ. Press, Cambridge.

Hagander, L. G., Midani, H. A., Kuskowski, M. A., and Parry, G. J. (2000). Quantitative sensory testing: effect of site and skin temperature on thermal thresholds. *Clinical Neurophysiology*, 111(1):17–22.

Hall, E. T. (1966). *The Hidden Dimension*. Doubleday.

Halvey, M., Henderson, M., Brewster, S. A., Wilson, G., and Hughes, S. A. (2012a). Augmenting media with thermal stimulation. In *Haptic and Audio Interaction Design, Conference Proceedings*, volume 7468 of *Lecture Notes in Computer Science*. Springer.

Halvey, M., Wilson, G., Brewster, S., and Hughes, S. (2012b). Baby it's cold outside: the influence of ambient temperature and humidity on thermal feedback. In *Human Factors In Computing Systems (CHI), Conference Proceedings*. ACM.

Halvey, M., Wilson, G., Brewster, S. A., and Hughes, S. A. (2013). Perception of thermal stimuli for continuous interaction. In *Human Factors in Computing Systems (CHI - Extended Abstracts), Conference Proceedings*. ACM.

Halvey, M., Wilson, G., Vazquez-Alvarez, Y., Brewster, S. A., and Hughes, S. A. (2011). The effect of clothing on thermal feedback perception. In *Multimodal Interfaces (ICMI), Conference Proceedings*. ACM.

Hamilton, C. (2011). *Communicating for results: A guide for business and the professions*. Thomson/Wadsworth, Belmont, Calif., 9th ed. edition.

Hardaway, R. A. (1990). Subliminally activated symbiotic fantasies: facts and artifacts. *Psychological bulletin*, 107(2):177–195.

Harrison, J. L. and Davis, K. D. (1999). Cold-evoked pain varies with skin type and cooling rate: a psychophysical study in humans. *Pain*, 83(2):123–135.

Hartmann, B., Mancini, M., and Pelachaud, C. (2005). Implementing expressive gesture synthesis for embodied conversational agents. In *Gesture in Human-Computer Interaction and Simulation, Workshop Proceedings*, volume 3881 of *Lecture Notes in Computer Science*, pages 188–199. Springer.

Hazlewood, W. R., Connelly, K., Makice, K., and Lim, Y.-K. (2008). Exploring evaluation methods for ambient information systems. In *Human Factors in Computing Systems (CHI - Extended Abstracts), Conference Proceedings*. ACM.

Heiner, J. M., Hudson, S. E., and Tanaka, K. (1999). The information percolator: ambient information display in a decorative object. In *User Interface Software and Technology (UIST), Conference Proceedings*. ACM.

Henze, N., Heuten, W., and Boll, S. (2006). Non-intrusive somatosensory navigation support for blind pedestrians. In *Eurohaptics, Conference Proceedings*.

Hertenstein, M. J., Keltner, D., App, B., Bulleit, B. A., and Jaskolka, A. R. (2006). Touch communicates distinct emotions. *Emotion (Washington, D.C.)*, 6(3):528–533.

Hess, U. and Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, 17(2):142–157.

Hill, J. W. and Bliss, J. C. (1968). Perception of sequentially presented tactile point stimuli. *Perception & Psychophysics*, 4(5):289–295.

Hollandsworth, J. G., Kazelskis, R., Stevens, J., and Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology*, 32(2):359–367.

Hong, J. (2013). Considering privacy issues in the context of google glass. *Communications of the ACM*, 56(11):10–11.

Hoque, M. A., Siekkinen, M., Khan, K. N., Xiao, Y., and Tarkoma, S. (2016). Modeling, profiling, and debugging the energy consumption of mobile devices. *ACM Computing Surveys (CSUR)*, 48(3):39.

Hoque, M. E., Courgeon, M., Martin, J., Mutlu, B., and Picard, R. W. (2013). Mach: My automated conversation coach. In *Pervasive and Ubiquitous Computing (UbiComp), Conference Proceedings*. ACM.

Horvitz, E. and Apacible, J. (2003). Learning and reasoning about interruption. In *Multimodal Interfaces (ICMI), Conference Proceedings*. ACM.

Hu, P., Shen, G., Jiang, X., Shih, S.-f., Lu, D., Zhao, F., Hong, D., Li, Q., Nirjon, S., Dickerson, R., and Stankovic, J. A. (2012). Septimu 2 - earphones for continuous and non-intrusive physiological and environmental monitoring. In *Embedded Network Sensor Systems (SenSys), Conference Proceedings*. ACM.

Hua, G., Yang, T.-Y., and Vasireddy, S. (2007). Peye: Toward a visual motion based perceptual interface for mobile devices. In *Human–Computer Interaction, Workshop Proceedings*, volume 4796 of *Lecture Notes in Computer Science*, pages 39–48. Springer.

Iacoboni, M. (2008). *Mirroring people: The new science of how we connect with others*. Farrar Straus and Giroux, New York, 1 edition.

IJzerman, H. and Semin, G. R. (2009). The thermometer of social relations mapping social proximity on temperature. *Psychological Science*, 20(10):1214–1220.

Iqbal, S. T. and Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *Human Factors in Computing Systems (CHI - Extended Abstracts), Conference Proceedings*, page 1489. ACM.

Ishimaru, S., Kunze, K., Kise, K., Weppner, J., Dengel, A., Lukowicz, P., and Bulling, A. (2014). In the blink of an eye: combining head motion and eye blink frequency for activity recognition with google glass. In *Augmented Human Conference*. ACM.

Iwata, H., Yano, H., Uemura, T., and Moriya, T. (2004). Food simulator: a haptic interface for biting. In *Virtual Reality (VR), Conference Proceedings*, pages 51–57. IEEE.

James, W. (1890). *The Principles Of Psychology*, volume 1. Macmillan.

Jiang, J., Coffey, P., and Toohey, B. (2006). Improvement of odor intensity measurement using dynamic olfactometry. *Journal of the Air & Waste Management Association*, 56(5):675–683.

Johansson, R. S. (1978). Tactile sensibility in the human hand: Receptive field characteristics of mechanoreceptive units in the glabrous skin area. *The Journal of Physiology*, 281(1):101–125.

Johnston, W. A., Hawley, K. J., Plewe, S. H., Elliott, J. M. G., and DeWitt, M. J. (1990). Attention capture by novel stimuli. *Journal of Experimental Psychology: General*, 119(4):397.

Jones, B. (1981). The developmental significance of cross-modal matching. In *Intersensory Perception and Sensory Integration*, pages 109–136. Springer US, Boston, MA.

Jones, B. and O'Neil, S. (1985). Combining vision and touch in texture perception. *Perception & Psychophysics*, 37(1):66–72.

Jones, H., Sabouret, N., Damian, I., Baur, T., André, E., Porayska-Pomsta, K., and Rizzo, P. (2014). Interpreting social cues to generate credible affective reactions of virtual job interviewers. In *Intelligent Digital Games for Empowerment (IDGEI), satelite Workshop to Intelligent User Interfaces (IUI), Workshop Proceedings*. ACM.

Jones, L. A. and Berris, M. (2002). The psychophysics of temperature perception and thermal-interface design. In *Haptic Interfaces for Virtual Environment and Teleoperator Systems, Conference Proceedings*, page 137. IEEE.

Jones, S. E. and Yarbrough, A. E. (1985). A naturalistic study of the meanings of touch. *Communication Monographs*, 52(1):19–56.

Jovanov, E., Milosevic, M., and Milenkovic, A. (2013). A mobile system for assessment of physiological response to posture transitions. In *Engineering in Medicine and Biology Society (EMBS), Conference Proceedings*, volume 2013, pages 7205–7208. IEEE.

Ju, W. (2015). *The Design of Implicit Interactions*. Morgan & Claypool.

Kaaresoja, T. and Linjama, J. (2005). Perception of short tactile pulses generated by a vibration motor in a mobile phone. In *Haptic Interfaces for Virtual Environment and Teleoperator Systems, Conference Proceedings*, pages 471–472. IEEE.

Kahneman, D. (1973). *Attention and effort*. Prentice Hall series in experimental psychology. Prentice Hall, Englewood Cliffs.

Kahneman, D., Peavler, W. S., and Onuska, L. (1968). Effects of verbalization and incentive on the pupil response to mental activity. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 22(3):186–196.

Kahneman, D., Tursky, B., Shapiro, D., and Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, 79(1p1):164.

Kanner, L. (1968). Autistic disturbances of affective contact. *Acta paedopsychiatrica*, 35(4):100–136.

Kappas, A. P. (2013). Social regulation of emotion: messy layers. *Frontiers in Psychology*, 4:51.

Kaye, J. N. (2001). *Symbolic Olfactory Display*. Master thesis, supervised by Hawley, M. J., Massachusetts Institute of Technology, Massachusetts, USA.

Keast, R. S. and Breslin, P. A. (2003). An overview of binary taste–taste interactions. *Food Quality and Preference*, 14(2):111–124.

Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*. CUP Archive.

Kenshalo, D. R., Holmes, C. E., and Wood, P. B. (1968). Warm and cool thresholds as a function of rate of stimulus temperature change. *Perception & Psychophysics*, 3(2):81–84.

Kistler, F., Endrass, B., Damian, I., Dang, C. T., and André, E. (2012). Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces*, 6(1):39–47.

Kline, T. J., Ghali, L. M., Kline, D. W., and Brown, S. (1990). Visibility distance of highway signs among young, middle-aged, and older observers: Icons are better than text. *Human Factors*, 32(5):609–619.

Kosmalla, F., Wiehr, F., Daiber, F., Krüger, A., and Löchtefeld, M. (2016). Climbaware: Investigating perception and acceptance of wearables in rock climbing. In *Human Factors in Computing Systems (CHI), Conference Proceedings*. ACM.

Koutropoulos, A., Keskin, N., Abajian, S. C., Hogue, R., Rodriguez, C. O., and Gallagher, M. S. (2011). Exploring the mooc format as a pedagogical approach for mlearning. In *Mobile and Contextual Learning, Conference Proceedings*, pages 138–145.

Kramer, G., Walker, B., and Bargar, R. (1999). *Sonification report: Status of the field and research agenda*. International Community for Auditory Display.

Kukkonen, J., Lagerspetz, E., Nurmi, P., and Andersson, M. (2009). BeTelGeuse: A platform for gathering and processing situational data. *IEEE Pervasive Computing*, 8(2):49–56.

Larkin, J. and Simon H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100.

Lathia, N., Rachuri, K., Mascolo, C., and Roussos, G. (2013). Open source smartphone libraries for computational social science. In *Pervasive and Ubiquitous Computing, Workshop Proceedings*. ACM.

Lawless, H. T. (1997). Olfactory psychophysics. In *Tasting and smelling*, Handbook of perception and cognition, 2nd edition. Academic Press, San Diego.

Lee, J.-H., Poliakoff, E., and Spence, C. (2009). The effect of multimodal feedback presented via a touch screen on the performance of older adults. In *Haptic and Audio Interaction Design (HAID), Conference Proceedings*. Springer.

Lee, W. and Lim, Y.-K. (2010). Thermo-message: exploring the potential of heat as a modality of peripheral expression. In *Human Factors in Computing Systems (CHI - Extended Abstracts), Conference Proceedings*. ACM.

Lee, W. and Lim, Y.-K. (2012). Explorative research on the heat as an expression medium: focused on interpersonal communication. *Personal and Ubiquitous Computing*, 16(8):1039–1049.

Lee, Y., Min, C., Hwang, C., Lee, J., Hwang, I., Ju, Y., Yoo, C., Moon, M., Lee, U., and Song, J. (2013). Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Mobile Systems, Applications and Services (MobiSys), Conference Proceedings*, pages 375–388. ACM.

Lettvin, J. Y. (1976). On seeing sidelong. *The Sciences*, 16(4):10–20.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Research and Development in Information Retrieval, Conference Proceedings*, pages 3–12. Springer.

LiKamWa, R., Liu, Y., Lane, N. D., and Zhong, L. (2013). Moodscope: building a mood sensor from smartphone usage patterns. In *Mobile Systems, Applications and Services (MobiSys), Conference Proceedings*. ACM.

Liu, J., Wang, C., Qiao, J., Wang, W., and Zhang, Y. (2012). A machine-to-machine application: Body posture recognition on smartphones for intelligent remote control. In *Cloud Computing and Intelligence Systems (CCIS), Conference Proceedings*, pages 884–888. IEEE.

Liu, J., Zhong, L., Wickramasuriya, J., and Vasudevan, V. (2009). uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6):657–675.

Liu, Y. C. (2001). Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveller information systems. *Ergonomics*, 44(4):425–442.

Loomis, J. M. (1981). Tactile pattern perception. *Perception*, 10(1):5–27.

Loomis, J. M. and Lederman, S. J. (1986). Tactual perception. In *Handbook of perception and human performances*. John Wiley & Sons.

Lu, H., Brush, A. J. B., Priyantha, B., Karlson, A. K., and Liu, J. (2011). Speakersense: Energy efficient unobtrusive speaker identification on mobile phones. In *Pervasive Computing, Conference Proceedings*, Pervasive'11, pages 188–205, Berlin, Heidelberg. Springer-Verlag.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. (2012). StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Ubiquitous Computing (UbiComp), Conference Proceedings*, pages 351–360. ACM.

Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., and Campbell, A. T. (2010). The jigsaw continuous sensing engine for mobile phone applications. In *Embedded Networked Sensor Systems (SenSys), Conference Proceedings*. ACM.

Lupton, E. (2004). *Thinking with Type: A Primer for Deisgners: A Critical Guide for Designers, Writers, Editors, & Students*. Princeton Architectural Press.

Madan, A. (2005). *Thin slices of interest: Thin slices of interest*. Phd thesis, supervised by Pentland, A., Massachusetts Institute of Technology.

Madan, A. and Pentland, A. (2006). Vibefones: Socially aware mobile phones. In *Wearable Computers, Conference Proceedings*, pages 109–112. IEEE.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., and Ames, M. (2003). Heuristic evaluation of ambient displays. In *Human Factors in Computing Systems (CHI), Conference Proceedings*. ACM.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15(2):197–237.

Matic, A., Osmani, V., Maxhuni, A., and Mayora, O. (2012). Multi-modal mobile sensing of social interactions. In *Pervasive Computing Technologies for Healthcare, Conference Proceedings*. IEEE.

McAtamney, G. and Parker, C. (2006). An examination of the effects of a wearable display on informal face-to-face communication. In *Conference on Human Factors in Computing Systems (CHI)*, pages 45–54. ACM.

McCrickard, D., Catrambone, R., Chewar, C., and Stasko, J. T. (2003). Establishing tradeoffs that leverage attention for utility: Empirically evaluating information display in notification systems. *International Journal of Human-Computer Studies*, 58(5):547–582.

McNaney, R., Poliakov, I., Vines, J., Balaam, M., Zhang, P., and Olivier, P. (2015). Lapp: A speech loudness application for people with parkinson's on google glass. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 497–500. ACM.

McNeill, D. (1992). Hand and mind: What gestures reveal about thoughts. *University of Chicago Press*.

Mehlmann, G., Janowski, K., Baur, T., Häring, M., André, E., and Gebhard, P. (2014). Modeling gaze mechanisms for grounding in hri. In *Artificial Intelligence (ECAI), Conference Proceedings*. IOS Press.

Mehrabian, A. (1969). Some referents and measures of nonverbal behavior. *Behavior Research Methods & Instrumentation*, 1(6):203–207.

Mehrabian, A. (1981). *Silent messages: Implicit Communication of Emotions and Attitudes*. Wadsworth Publishing Co Inc, Belmont.

Milewski, A. E. and Iaccino, J. (1982). Strategies in cross-modality matching. *Perception & Psychophysics*, 31(3):273–275.

Miller, E. A. (1972). Interaction of vision and touch in conflict and nonconflict form perception tasks. *Journal of Experimental Psychology*, 96(1):114–123.

Millodot, M. (2014). *Dictionary of Optometry and Visual Science*. Elsevier Health Sciences UK.

Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., and Campbell, A. T. (2008). Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Embedded Network Sensor Systems (SenSys), Conference Proceedings*, pages 337–350. ACM.

Miyata, Y. and Norman, D. A. (1986). Psychological issues in support of multiple activities. In *User Centered System Design; New Perspectives on Human-Computer Interaction*, pages 265–284. L. Erlbaum Associates Inc.

Monk, C. A., Boehm-Davis, D. A., and Trafton, J. G. (2004). Recovering from interruptions: implications for driver distraction research. *Human Factors*, 46(4):650–663.

Moray, N. (1967). Where is capacity limited? a survey and a model. *Acta Psychologica*, 27:84–92.

Moser, C. and Tscheligi, M. (2013). Playful taste interaction. In *Interaction Design and Children (IDC), Conference Proceedings*. ACM.

Muaremi, A., Arnrich, B., and Troster, G. (2013). Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience*, 3:172–183.

Müller, H., Kazakova, A., Pielot, M., Heuten, W., and Boll, S. (2013). Ambient timer – unobtrusively reminding users of upcoming tasks with ambient light. In *Human-Computer Interaction (INTERACT), Conference Proceedings*, pages 211–228. Springer.

Müller, H., Löcken, A., Heuten, W., and Boll, S. (2014). Sparkle: an ambient light display for dynamic off-screen points of interest. In *Human-Computer Interaction (NordiCHI), Conference Proceedings*. ACM.

Muralidhar, S., Costa, J. M. R., Nguyen, L. S., and Gatica-Perez, D. (2016). Dites-moi: wearable feedback on conversational behavior. In *Mobile and Ubiquitous Multimedia, Conference Proceedings*, pages 261–265. ACM.

Murer, M., Aslan, I., and Tscheligi, M. (2013). Lollio: Exploring taste as playful modality. In *Tangible, Embedded and Embodied Interaction (TEI), Conference Proceedings*, pages 299–302. ACM.

Mutlu, B., Forlizzi, J., and Hodgins, J. (2006). A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid Robots, Conference Proceedings*, pages 518–523. IEEE.

Myers, C. S. and Rabiner, L. R. (1981). A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7):1389–1409.

Nakamura, Y., Goto, T. K., Tokumori, K., Yoshiura, T., Kobayashi, K., Nakamura, Y., Honda, H., Ninomiya, Y., and Yoshiura, K. (2012). The temporal change in the cortical activations due to salty and sweet tastes in humans: fmri and time-intensity sensory evaluation. *Neuroreport*, 23(6):400–404.

Narasimhan, M., Viola, P., and Shilman, M. (2006). Online decoding of markov models under latency constraints. In *Machine Learnine, Conference Proceedings*, pages 657–664. ACM.

Navon, D. (1984). Resources–a theoretical soup stone? *Psychological Review*, 91(2):216–234.

Navon, D. and Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, 86(3):214–255.

Nesbitt, K. (2003). *Designing multi-sensory displays for abstract data*. Phd thesis, University of Sydney, Sydney, Australia.

Neukrug, E. S. (1991). Computer–assisted live supervision in counselor skills training. *Counselor Education and Supervision*, 31(2):132–138.

Nguyen, A.-T., Chen, W., and Rauterberg, M. (2012). Online feedback system for public speakers. In *E-Learning, e-Management and e-Services (IS3e), Conference Proceedings*, pages 1–5. IEEE.

Nirjon, S., Dickerson, R. F., Asare, P., Li, Q., Hong, D., Stankovic, J. A., Hu, P., Shen, G., and Jiang, X. (2013). Auditeur: a mobile-cloud service platform for acoustic event detection on smartphones. In *Mobile Systems, Applications and Services (MobiSys), Conference Proceedings*. ACM.

Notebaert, L., Crombez, G., van Damme, S., de Houwer, J., and Theeuwes, J. (2011). Signals of threat do not capture, but prioritize, attention: A conditioning approach. *Emotion*, 11(1):81.

Obaid, M., Damian, I., Kistler, F., Endrass, B., Wagner, J., and André, E. (2012). Cultural behaviors of virtual agents in an augmented reality environment. In *Intelligent Virtual Agents (IVA), Conference Publication*, volume 7502 of *Lecture Notes in Computer Science*, pages 412–418, Berlin, Heidelberg. Springer-Verlag.

Obrist, M., Comber, R., Subramanian, S., Piqueras-Fiszman, B., Velasco, C., and Spence, C. (2014). Temporal, affective, and embodied characteristics of taste experiences. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 2853–2862. ACM.

Occhialini, V., van Essen, H., and Eggen, B. (2011). Design and evaluation of an ambient display to support time management during meetings. In *Human-Computer Interaction (INTERACT), Conference Proceedings*, pages 263–280. Springer.

Ofek, E., Iqbal, S. T., and Strauss, K. (2013). Reducing disruption from subtle information delivery during a conversation: Mode and bandwidth investigation. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 3111–3120. ACM.

Okoshi, T., Tsubouchi, K., Taji, M., Ichikawa, T., and Tokuda, H. (2017). Attention and engagement-awareness in the wild: A large-scale study with adaptive notifications. In *Pervasive Computing and Communications, Conference Proceedings*. IEEE.

Olson, H. F. (1972). The measurement of loudness. *Audio Magazine*, pages 18–22.

Ouwerkerk, M., Pasveer, F., and Langereis, G. (2008). Unobtrusive sensing of psychophysiological parameters. In *Probing Experience*, volume 8 of *Philips Research*, pages 163–193. Springer Science + Business Media B. V, Dordrecht.

Oviatt, S. (2003). Multimodal interfaces. In *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, pages 286–304. CRC Press.

Oviatt, S., Coulston, R., and Lunsford, R. (2004). When do we interact multimodally?: cognitive load and multimodal communication patterns. In *Multimodal Interfaces (ICMI), Conference Proceedings*. ACM.

Palaghias, N., Hoseinitabatabaei, S. A., Nati, M., Gluhak, A., and Moessner, K. (2016). A survey on mobile social signal processing. *ACM Computing Surveys (CSUR)*, 48(4):57.

Pan, X., Gillies, M., Barker, C., Clark, D. M., and Slater, M. (2012). Socially anxious and confident men interact with a forward virtual woman: an experimental study. *PloS one*, 7(4):e32931.

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M., and Vinciarelli, A. (2011). Social signal processing: The research agenda. In *Visual Analysis of Humans*, pages 511–538. Springer.

Park, T., Lee, J., Hwang, I., Yoo, C., Nachman, L., and Song, J. (2011). E-gesture: a collaborative architecture for energy-efficient gesture recognition with hand-worn sensor and mobile devices. In *Embedded Networked Sensor Systems, Conference Proceedings*, pages 260–273. ACM.

Patil, V., Akhtar, M. Q., Parab, A., and Fernandes, A. (2012). Sonification of facial expression using dense optical flow on segmented facial plane. In *Computing and Control Engineering (ICCCE), Conference Proceedings*. Coimbatore Institute of Information Technology.

Pease, A. (1988). *Body Language: How to read other's thoughts by their gestures*. Sheldon Press, London, tenth impression 1988 edition.

Pease, B. and Pease, A. (2008). *The Definitive Book of Body Language*. Random House Publishing Group.

Peng, C., Shen, G., Zhang, Y., Li, Y., and Tan, K. (2007). Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Embedded Networked Sensor Systems (SensSys), Conference Proceedings*, pages 1–14. ACM.

Pentland, A. (2007). Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108–111.

Pertovaara, A. and Kojo, I. (1985). Influence of the rate of temperature change on thermal thresholds in man. *Experimental neurology*, 87(3):439–445.

Pfaffmann, C. (1980). Wundt's schema of sensory affect in the light of research on gustatory preferences. *Psychological Research*, 42(1-2):165–174.

Pham, P. and Wang, J. (2016). Adaptive review for mobile mooc learning via implicit physiological signal sensing. In *Multimodal Interaction (ICMI), Conference Proceedings*, pages 37–44. ACM.

Pick, H. L., Warren, D. H., and Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, 6(4):203–205.

Pielot, M., Poppinga, B., and Boll, S. (2010). Pocketnavigator: vibro-tactile waypoint navigation for everyday mobile devices. In *Human Computer Interaction with Mobile Devices and Services (MobileHCI), Conference Proceedings*. ACM.

Pielot, M., Poppinga, B., Heuten, W., and Boll, S. (2011). A tactile compass for eyes-free pedestrian navigation. In *Human-Computer Interaction (INTERACT), Conference Proceedings*, volume 6947, pages 640–656. Springer.

Pierce, C. S. and Jastrow, J. (1984). On small differences in sensation. In *Memoirs of the National Academy of Sciences*, pages 73–83. Government Printing Office.

Pineau, N., Schlich, P., Cordelle, S., Mathonnière, C., Issanchou, S., Imbert, A., Rogeaux, M., Etiévant, P., and Köster, E. (2009). Temporal dominance of sensations: Construction of the tds curves and comparison with time–intensity. *Food Quality and Preference*, 20(6):450–455.

Pino, C. and Kavasidis, I. (2012). Improving mobile device interaction by eye tracking analysis. In *Computer Science and Information Systems, Conference Proceedings*. IEEE.

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., Al'absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., Smailagic, A., and Wittmers, L. E. (2011). Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Information Processing in Sensor Networks, Conference Proceedings*. IEEE.

Poh, M.-Z., Swenson, N. C., and Picard, R. W. (2010). Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):786–794.

Politis, I., Brewster, S., and Pollick, F. (2013). Evaluating multimodal driver displays of varying urgency. In *Automotive User Interfaces and Interactive Vehicular Applications, Conference Proceedings*. ACM.

Poppinga, B., Heuten, W., and Boll, S. (2014). Sensor-based identification of opportune moments for triggering notifications. *IEEE Pervasive Computing*, 13(1):22–29.

Porayska-Pomsta, K., Anderson, K., Damian, I., Baur, T., André, E., Bernardini, S., and Rizzo, P. (2013). Modelling users' affect in job interviews: Technological demo. In *User Modeling, Adaptation, and Personalization*, volume 7899, pages 353–355. Springer.

Porayska-Pomsta, K., Rizzo, P., Damian, I., Baur, T., André, E., Sabouret, N., Jones, H., Anderson, K., and Chryssafidou, E. (2014). Who's afraid of job interviews? definitely a question for user modelling. In *User Modeling, Adaptation, and Personalization (UMAP), Conference Proceedings*, volume 8538 of *Lecture Notes in Computer Science*, pages 411–422. Springer.

Posthuma, R. A., Morgeson, F. P., and Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Journal of Personnel Psychology*, 55(1):1–81.

Pousman, Z. and Stasko, J. (2006). A taxonomy of ambient information systems: four patterns of design. In *Advanced Visual Interfaces, Conference Proceedings*, pages 67–74. ACM.

Rachuri, K. K., Mascolo, C., Musolesi, M., and Rentfrow, P. J. (2011). Sociablesense: exploring the trade-offs of adaptive sampling and computation offloading for social sensing. In *Mobile Computing and Networking, Conference Proceedings*. ACM.

Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. (2010). Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Pervasive and Ubiquitous Computing, Conference Proceedings*. ACM.

Ramig, L., Sapir, S., Countryman, S., Pawlas, A., O'Brien, C., Hoehn, M., and Thompson, L. (2001). Intensive voice treatment (LSVT) for patients with parkinson's disease: a 2 year follow up. *Journal of Neurology, Neurosurgery, and Psychiatry*, 71(4):493–498.

Ranasinghe, N., Nakatsu, R., Nii, H., and Gopalakrishnakone, P. (2012). Tongue mounted interface for digitally actuating the sense of taste. In *Wearable Computers (ISWC), Conference Proceedings*, pages 80–87. IEEE.

Ranasinghe, N., Suthokumar, G., Lee, K.-Y., and Do, E. Y.-L. (2015). Digital flavor: Towards digitally simulating virtual flavors. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Raudenbush, B., Grayhem, R., Sears, T., and Eshun-Wilson, I. (2009). Effects of peppermint and cinnamon odor administration on simulated driving alertness, mood and workload. *North American Journal of Psychology*, (11(2)):245–256.

Remington, R. W., Johnston, J. C., and Yantis, S. (1992). Involuntary attentional capture by abrupt onsets. *Perception & Psychophysics*, 51(3):279–290.

Richmond, V. P. and McCroskey, J. C. (1998). *Communication: Apprehension, avoidance, and effectiveness*. Allyn and Bacon, Boston, 5th ed. edition.

Rock, I. and Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science*, 143(3606):594–596.

Ruf, T., Ernst, A., and Küblbeck, C. (2011). Face detection with the sophisticated high-speed object recognition engine (shore). *Microelectronic Systems*, pages 243–252.

Running, C. A., Craig, B. A., and Mattes, R. D. (2015). Oleogustus: The unique taste of fat. *Chemical Senses*.

Salminen, K., Surakka, V., Raisamo, J., Lylykangas, J., Pystynen, J., Raisamo, R., Mäkelä, K., and Ahmaniemi, T. (2011). Emotional responses to thermal stimuli. In *Multimodal Interfaces (ICMI), Conference Proceedings*. ACM.

Salvucci, D. D., Taatgen, N. A., and Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: from concurrent performance to task switching, interruption, and resumption. In *Human Factors in Computing Systems (CHI), Conference Proceedings*. ACM.

Sapouna, M., Wolke, D., Vannini, N., Watson, S., Woods, S., Schneider, W., Enz, S., Hall, L., Paiva, A., André, E., Dautenhahn, K., and Aylett, R. (2010). Virtual learning intervention to reduce bullying victimization in primary school: a controlled trial. *Journal of child psychology and psychiatry, and allied disciplines*, 51(1):104–112.

Schacter, D. L., Gilbert, D. T., and Wegner, D. M. (2011). *Psychology*. Worth Publishers, 2 edition.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2):227–256.

Scherl, C. R. and Haley, J. (2000). Computer monitor supervision: A clinical note. *The American journal of family therapy*, 28(3):275–282.

Schneider, J., Börner, D., van Rosmalen, P., and Specht, M. (2015). Presentation trainer, your public speaking multimodal coach. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Schneider, W. and Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science*, 27(3):525–559.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological Review*, 84(1):1.

Seizova-Cajić, T. (1998). Size perception by vision and kinesthesia. *Perception & Psychophysics*, 60(4):705–718.

Settles, B. (2012). *Active Learning*. Morgan & Claypool Publishers.

Shettleworth, S. J. (2009). *Cognition, Evolution, and Behavior*. Oxford University Press.

Sieverding, M. (2009). 'be cool!': Emotional costs of hiding feelings in a job interview. *International Selection and Assessment*, 17(4):391–401.

Sigman, M., Spence, S. J., and Wang, A. T. (2006). Autism from developmental and neuropsychological perspectives. *Annual review of clinical psychology*, 2:327–355.

Sigrist, R., Rauter, G., Riener, R., and Wolf, P. (2012). Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review*, 20(1):21–53.

Silverman, L. H. and Weinberger, J. (1985). Mommy and i are one: Implications for psychotherapy. *American Psychologist*, 40(12):1296.

Six, J., Cornelis, O., and Leman, M. (2014). Tarsosdsp, a real-time audio processing framework in java. In *Semantic Audio, Conference Proceedings*. AES.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis*. Appleton-Century-Crofts.

Smith, D. V. (1971). Taste intensity as a function of area and concentration. *Journal of Experimental Psychology*, 87(2):163–171.

Spehr, M., Kelliher, K. R., Li, X.-H., Boehm, T., Leinders-Zufall, T., and Zufall, F. (2006). Essential role of the main olfactory system in social recognition of major histocompatibility complex peptide ligands. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(7):1961–1970.

Stapleton, A. J. and Taylor, P. C. (2003). Why videogames are cool & school sucks! In *Australian Game Developers Conference (AGDC), Conference Proceedings*, volume 20, page 23.

Stevens, J. C. (1991). Thermal sensibility. *The psychology of touch*, pages 61–90.

Stevens, J. C., Cain, W. S., and Burke, R. J. (1988). Variability of olfactory thresholds. *Chemical Senses*, 13(4):643–653.

Stevens, S. S. (1955). The measurement of loudness. *The Journal of the Acoustical Society of America*, 27(5):815–829.

Strandvall, T. (2009). Eye tracking in human-computer interaction and usability research. In *Human-Computer Interaction (INTERACT), Conference Proceedings*, volume 5727 of *Lecture Notes in Computer Science*, pages 936–937. Springer.

Strasburger, H., Rentschler, I., and Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13.

Suhonen, K., Müller, S., Rantala, J., Väänänen-Vainio-Mattila, K., Raisamo, R., and Lantz, V. (2012). Haptically augmented remote speech communication: a study of user practices and experiences. In *Human-Computer Interaction (NordiCHI), Conference Proceedings*. ACM.

Szczerba, J., Hersberger, R., and Riegelman, A. (2012). Design and evaluation of a differential speedometer. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):1629–1633.

Tan, H. Z., Durlach, N. I., Reed, C. M., and Rabinowitz, W. M. (1999). Information transmission with a multifinger tactual display. *Perception & Psychophysics*, 61(6):993–1008.

Tanveer, M. I., Lin, E., and Hoque, M. (2015). Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Intelligent User Interfaces (IUI), Conference Proceedings*, pages 286–295. ACM.

Tanveer, M. I., Zhao, R., Chen, K., Tiet, Z., and Hoque, M. E. (2016). Automanner: An automated interface for making public speakers aware of their mannerisms. In *Intelligent User Interfaces (IUI), Conference Proceedings*. ACM.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta Psychologica*, 135(2):77–99.

Ting-Toomey, S. (1999). Communicating across cultures. *The Guilford Press*.

Trumbo, D. and Noble, M. (1970). Secondary task effects on serial verbal learning. *Journal of Experimental Psychology*, 85(3):418.

Turin, L. (1996). A spectroscopic mechanism for primary olfactory reception. *Chemical Senses*, 21(6):773–791.

Twardon, L., Koesling, H., Finke, A., and Ritter, H. (2013). Gaze-contingent audio-visual substitution for the blind and visually impaired. In *Pervasive Computing Technologies for Healthcare, Conference Proceedings*. ICST.

van der Linden, J., Johnson, R., Bird, J., Rogers, Y., and Schoonderwaldt, E. (2011a). Buzzing to play: Lessons learned from an in the wild study of real-time vibrotactile feedback. In *Human Factors in Computing Systems (CHI), Conference Proceedings*, pages 533–542. ACM.

van der Linden, J., Schoonderwaldt, E., Bird, J., and Johnson, R. (2011b). Musicjacket—combining motion capture and vibrotactile feedback to teach violin bowing. *IEEE Transactions on Instrumentation and Measurement*, 60(1):104–113.

Verrillo, R. T., Fraioli, A. J., and Smith, R. L. (1969). Sensation magnitude of vibrotactile stimuli. *Perception & Psychophysics*, 6(6):366–372.

Vidal, M., Nguyen, D. H., and Lyons, K. (2014). Looking at or through? In *Wearable Computers (ISWC), Conference Proceedings*, pages 87–90. ACM.

Vinciarelli, A., Murray-Smith, R., and Bourlard, H. (2010). Mobile social signal processing: vision and research issues. In *Human Computer Interaction with Mobile Devices and Services (MobileHCI), Conference Proceedings*. ACM.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Vinciarelli, A., Pantic, M., Bourlard, H., and Pentland, A. (2008). Social signal processing. In *Multimedia (MM), Conference Proceedings*, page 1061. ACM.

Wagner, J. (2016). *Social Signal Interpretation: Building Online Systems for Multimodal Behaviour Analysis*. Phd thesis, supervised by André, E., Universität Augsburg, Augsburg.

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., and André, E. (2013). The social signal interpretation (SSI) framework - multimodal signal processing and recognition in real-time. In *Multimedia (MM), Conference Proceedings*, Barcelona.

Wallace, F., Flanery, J., and Knezek, G. A. (1991). The effect of subliminal help presentations on learning a text editor. *Information Processing & Management*, 27(2-3):211–218.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896.

Wang, Y., Lin, J., Annavaram, M., Jacobson, Q. A., Hong, J., Krishnamachari, B., and Sadeh, N. (2009). A framework of energy efficient mobile sensing for automatic user state recognition. In *Mobile Systems (MobiSys), Conference Proceedings*. ACM.

Warnock, D., McGee-Lennon, M., and Brewster, S. (2011). The role of modality in notification performance. In *Human-Computer Interaction (INTERACT), Conference Proceedings*, pages 572–588. Springer.

Warren, D. H. and Cleaves, W. T. (1971). Visual-proprioceptive interaction under large amounts of conflict. *Journal of Experimental Psychology*, 90(2):206–214.

Weinberger, J. (1992). Validating and demystifying subliminal psychodynamic activation. In *Perception without awareness: Cognitive, clinical, and social perspectives*, pages 170–188. Guilford Press.

Weinstein, S. (1968). Intensive and extensive aspects of tactile sensitivity as a function of body part, sex and laterality. In *The Skin Senses, Charles C. Thomas*. Springfield.

Weiser, M. and Brown, J. (1995). Designing calm technology. *Powergrid Journal*.

Wickens, C., Isreal, J., and Donchin, E. (1977). The event-related cortical potential as an index of task workload. In *Human Factors Society, Meeting Proceedings*.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):449–455.

Wickens, C. D. and McCarley, J. S. (2007). *Applied Attention Theory*. CRC Press.

Willemse, C. J., Munters, G. M., van Erp, J. B., and Heylen, D. (2015). Nakama: A companion for non-verbal affective communication. In *Multimodal Interaction (ICMI), Conference Proceedings*. ACM.

Williams, L. E. and Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322(5901):606–607.

Wilson, G. and Brewster, S. A. (2017). Multi-moji: Combining thermal, vibrotactile & visual stimuli to expand the affective range of feedback. In *Human Factors in Computing Systems, Conference Proceedings*. ACM.

Wilson, G., Davidson, G., and Brewster, S. A. (2015). In the heat of the moment: Subjective interpretations of thermal feedback during interaction. In *Human Factors in Computing Systems (CHI), Conference Proceedings*. ACM.

Wilson, G., Halvey, M., Brewster, S. A., and Hughes, S. A. (2011). Some like it hot: thermal feedback for mobile devices. In *Human Factors in Computing Systems (CHI), Conference Proceedings*. ACM.

Wooten, B. R. and Wald, G. (1973). Color-vision mechanisms in the peripheral retinas of normal and dichromatic observers. *The Journal of general physiology*, 61(2):125–145.

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., and Freeman, W. T. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics - Proceedings SIGGRAPH*, (31).

Xia, C. and Maes, P. (2013). The design of artifacts for augmenting intellect. In *Augmented Human (AH), Conference Proceedings*. ACM.

Xu, B., Yu, R., Sun, G., and Yang, Z. (2011). Whistle: Synchronization-free tdoa for localization. In *Distributed Computing Systems, Conference Proceedings*, pages 760–769. IEEE.

Yang, X., You, C.-W., Lu, H., Lin, M., Lane, N. D., and Campbell, A. T. (2012). Visage: A face interpretation engine for smartphone applications. In *Mobile Computing, Applications, and Services, Conference Proceedings*, pages 149–168. Springer.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, volume 6, pages 567–577.

Yoshida, M., Kato, C., Kawasumi, M., Yamasaki, H., Yamamoto, S., Nakano, T., and Yamada, M. (2011). Study on stimulation effects for driver based on fragrance presentation. In *Machine Vision Applications, Conference Proceedings*, pages 332–335. IAPR.

Zehetleitner, M., Koch, A. I., Goschy, H., and Müller, H. J. (2013). Salience-based selection: Attentional capture by distractors less salient than the target. *PLoS One*, 8(1).

Zhao, J., Al-Aidroos, N., and Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychological science*, 24(5):667–677.

Zimmermann, P. and Fimm, B. (2004). A test battery for attentional performance. In *Applied neuropsychology of attention. Theory, diagnosis and rehabilitation*, pages 110–151. Psychology Press.

Appendix

A Feedback Strategy XML Schema

The following code block represents the XML schema definition (XSD) for SSJ strategy files. The file is also hosted online.[2]

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns="hcm.ssj" xmlns:xs="http://www.w3.org/2001/XMLSchema"
      targetNamespace="hcm.ssj" elementFormDefault="qualified" >

 <!--Type definitions-->
 <xs:simpleType name="valence">
  <xs:restriction base="xs:NMTOKEN">
   <xs:enumeration value="Desirable" />
   <xs:enumeration value="Undesirable" />
  </xs:restriction>
 </xs:simpleType>
 <xs:simpleType name="fb_type">
  <xs:restriction base="xs:NMTOKEN">
   <xs:enumeration value="visual" />
   <xs:enumeration value="audio" />
   <xs:enumeration value="tactile" />
  </xs:restriction>
 </xs:simpleType>
 <xs:simpleType name="fb_device">
  <xs:restriction base="xs:NMTOKEN">
   <xs:enumeration value="Myo" />
   <xs:enumeration value="MsBand" />
  </xs:restriction>
 </xs:simpleType>
 <xs:simpleType name="vibration_type">
  <xs:restriction base="xs:NMTOKEN">
   <xs:enumeration value="NOTIFICATION_ONE_TONE" />
   <xs:enumeration value="NOTIFICATION_TWO_TONE" />
   <xs:enumeration value="NOTIFICATION_ALARM" />
   <xs:enumeration value="NOTIFICATION_TIMER" />
   <xs:enumeration value="ONE_TONE_HIGH" />
   <xs:enumeration value="TWO_TONE_HIGH" />
   <xs:enumeration value="THREE_TONE_HIGH" />
   <xs:enumeration value="RAMP_UP" />
   <xs:enumeration value="RAMP_DOWN" />
  </xs:restriction>
```

[2] http://hcmlab.github.io/ssj/res/feedback.xsd

```
 </xs:simpleType>

 <!--The base element-->
 <xs:element name="ssj">
  <xs:complexType>
   <xs:sequence maxOccurs="unbounded">
    <xs:element ref="strategy" minOccurs="1" maxOccurs="1"/>
   </xs:sequence>
  </xs:complexType>
 </xs:element>

 <!--Strategy-->
 <xs:element name="strategy">
  <xs:complexType>
   <xs:sequence>
    <xs:element maxOccurs="unbounded" name="feedback">
     <xs:complexType>
      <xs:sequence>
       <xs:element name="condition" minOccurs="1" maxOccurs="1">
        <xs:complexType>
         <xs:attribute name="event" type="xs:string" use="required" />
         <xs:attribute name="sender" type="xs:string" use="required" />
         <xs:attribute name="type" type="xs:string" use="optional" />
         <xs:attribute name="history" type="xs:unsignedByte" use="optional" />
         <xs:attribute name="sum" type="xs:boolean" use="optional" />
         <xs:attribute name="from" type="xs:decimal" use="required" />
         <xs:attribute name="to" type="xs:decimal" use="optional" />
        </xs:complexType>
       </xs:element>
       <xs:element name="action" minOccurs="1" maxOccurs="1">
        <xs:complexType>
         <xs:attribute name="res" type="xs:string" use="optional" />
         <xs:attribute name="lock" type="xs:integer" use="optional" />
         <xs:attribute name="lockSelf" type="xs:integer" use="optional" />
         <xs:attribute name="intensity" type="xs:string" use="optional" />
         <xs:attribute name="duration" type="xs:string" use="optional" />
         <xs:attribute name="brightness" type="xs:float" use="optional" />
         <xs:attribute name="type" type="vibration_type" use="optional" />
        </xs:complexType>
       </xs:element>
      </xs:sequence>
      <xs:attribute name="type" type="fb_type" use="required" />
      <xs:attribute name="valence" type="valence" use="optional"/>
      <xs:attribute name="level" type="xs:short" use="optional" />
      <xs:attribute name="layout" type="xs:string" use="optional" />
      <xs:attribute name="position" type="xs:short" use="optional" />
      <xs:attribute name="fade" type="xs:integer" use="optional" />
      <xs:attribute name="def_brightness" type="xs:float" use="optional" />
      <xs:attribute name="device" type="fb_device" use="optional" />
     </xs:complexType>
    </xs:element>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
</xs:schema>
```

B Sensors Supported by SSJ

Sensor Name	Connection	Supported Data Streams
Android Inertial Sensor Array	internal	accelerometer, ambient temperature, game rotation vector, geomagnetic rotation vector, gravity, gyroscope, light, linear acceleration, magnetic field, orientation, pressure, proximity, relative humidity, rotation vector, pedometer (step counter), temperature, heart rate
Android Camera	internal	RGB video
Android Microphone	internal[1]	audio
Android GPS	internal	GPS
Angel Sensor	Bluetooth	BVP
Generic BLE Sensor	Bluetooth	any data compatible with Bluetooth LE standard
Estimote Beacon	Bluetooth	distance to beacon (in meters)
Empatica E3/4	Bluetooth	acceleration, battery, BVP, EDA (also known as GSR, SC)
Google Glass	internal	distance to eye (infrared sensor), blinking
Microsoft Band 1/2	Bluetooth	acceleration, altimeter, barometer, light, calories, walking distance, EDA (also known as GSR, SC), gyroscope, heart rate, IBI, pedometer (step counter), skin temperature
Myo	Bluetooth	acceleration, linear acceleration, EMG
Pressure Mat[2]	Bluetooth	pressure

[1] any microphone connected over wire or Bluetooth with an Android device is supported

[2] developed by Zhixin Huang and Prof. Jingyuan Cheng, Wearable Computing Lab, Technische Universität Braunschweig

Abbreviations: Bluetooth Low Energy (BLE), Blood-Volume-Pulse (BVP), Electro-Dermal-Activity (EDA), Galvanic-Skin-Response (GSR), Skin Conductance (SC), Inter-Beat-Interval (IBI), Electromyography (EMG)

C Output Devices Supported by SSJ

Device Name	Modality	Available configuration options
Android Device (e.g. Google Glass)	visual	timing, prominence, duration, scope[1], level of detail[1]
Headphones[2]	audio	timing, prominence, duration, scope[1], level of detail[1]
Myo	tactile (vibrations)	timing, prominence, duration, level of detail
Microsoft Band 2	tactile (vibrations)	timing, prominence, duration, level of detail

[1] indirectly by choosing a custom resource

[2] any headphones or speakers connected over wire or Bluetooth with an Android device

D Implemented SSJ Components

Sensors and Sensor Channels

Name	Type	Description
AndroidSensor	Sensor	Handles connection with standard Android sensors. Supported sensor types: accelerometer, ambient temperature, game rotation vector, geomagnetic rotation vector, gravity, gyroscope, light, linear acceleration, magnetic field, orientation, pressure, proximity, relative humidity, rotation vector, pedometer (step counter), temperature, heart rate
AndroidSensorChannel	SensorChannel	Pushes data from AndroidSensor into pipeline
AngelSensor	Sensor	Handles connection with the Angel armband.
BVPAngelChannel	SensorChannel	Pushes BVP data from Angel armband into pipeline
Microphone	Sensor	Handles connection with the Android microphone.
AudioChannel	SensorChannel	Pushes audio data into pipeline
BLESensor	Sensor	Handles generic BLE connections.
BVPAndisChannel	SensorChannel	Pushes BVP data from custom wearable into pipeline
BVPBLEChannel	SensorChannel	Pushes BVP data from generic BLE sensor into pipeline
CameraSensor	Sensor	Handles connection to Android camera.
CameraChannel	SensorChannel	Pushes video data from camera into pipeline
Empatica	Sensor	Handles connection to Empatica E3/E4 armband.
AccelerationChannel	SensorChannel	Pushes acceleration data from Empatica into pipeline
BatteryChannel	SensorChannel	Pushes Empatica's battery statistics into pipeline
BVPChannel	SensorChannel	Pushes BVP data from Empatica into pipeline
GSRChannel	SensorChannel	Pushes GSR data from Empatica into pipeline
IBIChannel	SensorChannel	Pushes IBI data from Empatica into pipeline
TemperatureChannel	SensorChannel	Pushes skin temperature data from Empatica into pipeline
EstimoteBeacon	Sensor	Handles connection to Estimote beacons.
BeaconChannel	SensorChannel	Pushes distance data (in meter) from a beacon into the pipeline
FileReader	Sensor	Reads a file from the SD card.
FileReaderChannel	SensorChannel	Pushes data read from file into pipeline
InfraredSensor	Sensor	Handles connection to Google Glass infrared sensor. Requires Google Glass to be rooted and the permissions for the infrared sensor to be set.
InfraredChannel	SensorChannel	Pushes infrared signal intensity into pipeline
GPSSensor	Sensor	Handles connection to Android GPS service.
GPSChannel	SensorChannel	Pushes GPS data into pipeline
BluetoothReader	Sensor	Handles Bluetooth connection to external device.
BluetoothChannel	SensorChannel	Pushes data received over Bluetooth into pipeline
SocketReader	Sensor	Handles network connection (e.g. WiFi) to external device.
SocketChannel	SensorChannel	Pushes data received over network into pipeline
MSBand	Sensor	Handles connection to Microsoft Band 1/2 armband.
AccelerationChannel	SensorChannel	Pushes acceleration data from armband into pipeline
BarometerChannel	SensorChannel	Pushes air pressure data from armband into pipeline
BrightnessChannel	SensorChannel	Pushes brightness data from the armband's ambient light sensor into pipeline
CaloriesChannel	SensorChannel	Pushes the total number of calories burned (since armband factory-reset) into pipeline
DistanceChannel	SensorChannel	Pushes travel distance (cm), current speed (cm/s), current pace (ms/m) and pedometer mode into pipeline
GSRChannel	SensorChannel	Pushes GSR data (in kOhms) from armband into pipeline

GyroscopeChannel	SensorChannel	Pushes gyroscope data from armband into pipeline
HeartRateChannel	SensorChannel	Pushes heart rate data from armband into pipeline
IBIChannel	SensorChannel	Pushes IBI data from armband into pipeline
PedometerChannel	SensorChannel	Pushes total number of steps (since armband factory-reset) into pipeline
SkinTempChannel	SensorChannel	Pushes skin temperature data from armband into pipeline
Myo	Sensor	Handles connection to Myo armband.
AccelerationChannel	SensorChannel	Pushes acceleration data from Myo into pipeline
DynAccelerationChannel	SensorChannel	Pushes dynamic (linear) acceleration data into pipeline
EMGChannel	SensorChannel	Pushes EMG data from Myo into pipeline
BluetoothPressureSensor	Sensor	Handles connection to custom pressure mat.
BluetoothPressure MatChannel	SensorChannel	Pushes pressure data into pipeline
Profiler	Sensor	Handles device profiling.
CPULoadChannel	SensorChannel	Pushes CPU load data into pipeline

Transformer

Name	Type	Description
AudioConvert	Transformer	Converts audio stream between short and float values.
Energy	Transformer	Computes energy of audio stream (uses TarsosDSP [Six et al., 2014] library).
Pitch	Transformer	Computes pitch of audio stream (uses TarsosDSP [Six et al., 2014] library).
Intensity	Transformer	Computes intensity of audio stream (uses PRAAT [de Jong and Wempe, 2009] toolbox).
GSRArousalEstimation	Transformer	Computes arousal from GSR stream.
AccelerationFeatures	Transformer	Computes various features from acceleration stream.
OverallActivation	Transformer	Computes overall activation (expressivity feature [Baur et al., 2013b]) from acceleration stream.
BlinkDetection	Transformer	Performs blink detection on Google Glass' infrared stream [Ishimaru et al., 2014].
ClassifierT	Transformer	Classifies a data stream according to a predefined model. Supported models: SVM, Naïve Bayes, ANN (using TensorFlow).
AvgVar	Transformer	Computes the average and variance of the input window.
Median	Transformer	Computes the median of the input window.
MinMax	Transformer	Computes the minimum and maximum of the input window.
Butfilt	Transformer	Applies a butterworth filter on the input stream.
Count	Transformer	Computes the number of samples from the input signal which are above zero.
Derivative	Transformer	Computes the first, second, third and fourth derivative of the input stream.
Envelope	Transformer	Computes the envelope of a stream (filter).
FFTFeat	Transformer	Computes the FFT coefficients of the input stream.
Functionals	Transformer	Computes various statistical features from the input stream: mean, energy, standard deviation, minimum, maximum, range, position of minimum, position of maximum, number of zeroes, number of peaks, frame length, signal path length
IIR	Transformer	Applies a second-order infinite impulse response filter on the input stream.
Invert	Transformer	Inverts the input stream.

Merge	Transformer	Merges two or more streams into one stream with multiple dimensions.
MvgAvgVar	Transformer	Computes the moving/sliding average and variance of the input stream.
MvgMinMax	Transformer	Computes the moving/sliding minimum and maximum of the input stream.
MvgMinMax	Transformer	Performs a moving normalization of the input stream.
PSD	Transformer	Computes the power spectral density of the input stream.
Selector	Transformer	Selects a specific dimensions from the input stream.
Spectogram	Transformer	Computes spectral value for specified frequency bands.
HRVSpectral	Transformer	Computes various spectral heart rate features.

Consumer

Name	Type	Description
AudioWriter	Consumer	Writes an audio stream to the SD card using the MP4 format.
WavWriter	Consumer	Writes an audio stream to the SD card using the WAV format.
CameraWriter	Consumer	Writes a video stream to the SD card using the MP4 format.
FileWriter	Consumer	Writes a generic stream to the SD card.
BluetoothWriter	Consumer	Transmits a stream over Bluetooth to another device.
SocketWriter	Consumer	Transmits a stream over network (e.g. WiFi) to another device.
MobileSSIConsumer	Consumer	Pushes data to an SSI pipeline.
SpeechRate	Consumer	Computes the speech rate of the input signal. Result is sent out as events.
ValueEventSender	Consumer	Sends out input stream as events.
ThresholdEventSender	Consumer	Sends out an event whenever input stream exceeds a predefined threshold.
Classifier	Consumer	Classifies a data stream according to a predefined model. Supported models: SVM, Naïve Bayes, ANN (using TensorFlow).
SignalPainter	Consumer	Renders an input stream as a graph.
CameraPainter	Consumer	Renders a video stream.
Logger	Consumer	Prints input stream to the console.

Event Handler

Name	Type	Description
FileEventWriter	EventHandler	Writes incoming events to the SD card.
BluetoothEventWriter	EventHandler	Transmits incoming events over Bluetooth to another device.
SocketEventWriter	EventHandler	Transmits incoming events over network (e.g. WiFi) to another device.
BluetoothEventWriter	EventHandler	Receives events using a Bluetooth connection from another device.
SocketEventWriter	EventHandler	Receives events using a network connection (e.g. WiFi) from another device.
EventPainter	EventHandler	Renders incoming events as a graph.
EventLogger	EventHandler	Prints incoming events to the console.
FeedbackManager	EventHandler	Manages the delivery of live feedback in response to behaviour events. Supports multiple output devices spanning three modalities: visual, auditory and tactile.

E Training a Model with SSI

```
//names of folders where data is stored
string users[] = { "01","02","03" };
string path = "C:/data/";

// define sample lists
SampleList gsrSamples, hrSamples, tempSamples;

// iterate through all users
for (string user : users)
{
  // read anno
  Annotation anno;
  string annoFile = path + user + "/anno.anno";
  ModelTools::LoadAnnotation(anno, annoFile.c_str());
  // split anno into smaller chunks
  Annotation anno_online;
  ModelTools::ConvertToContinuousAnnotation(anno, anno_online, 10);

  // read data streams
  ssi_stream_t gsrStream;
  string gsrFile = path + user + "/GSR.stream";
  FileTools::ReadStreamFile(gsrFile.c_str(), gsrStream);

  ssi_stream_t hrStream;
  string hrFile = path + user + "/HeartRate.stream";
  FileTools::ReadStreamFile(hrFile.c_str(), hrStream);

  ssi_stream_t tempStream;
  string tempFile = path + user + "/SkinTemperature.stream";
  FileTools::ReadStreamFile(tempFile.c_str(), tempStream);

  // convert data streams into sample lists
  // each sample list aggregates the data from all users
  ModelTools::LoadSampleList(gsrSamples, gsrStream, anno_online, user);
  ModelTools::LoadSampleList(hrSamples, hrStream, anno_online, user);
  ModelTools::LoadSampleList(tempSamples, tempStream, anno_online, user);
}

// extract features from each sample list
SampleList gsrFeatures;
Functionals func = ssi_create(Functionals, 0, true);
ModelTools::TransformSampleList(gsrSamples, gsrFeatures, *func);

SampleList hrFeatures;
func = ssi_create(Functionals, 0, true);
ModelTools::TransformSampleList(hrSamples, hrFeatures, *func);

SampleList tempFeatures;
func = ssi_create(Functionals, 0, true);
ModelTools::TransformSampleList(tempSamples, tempFeatures, *func);

// merge sample lists
SampleList allFeatures;
SampleList* samplelistArray[] = {gsrFeatures, hrFeatures, tempFeatures};
```

```
ModelTools::MergeSampleList(allFeatures, 3, samplelistArray);

// train an SVM model
SVM *model = ssi_create(SVM, 0, true);
Trainer trainer(model);
trainer.train(allFeatures);

// save it to a file
trainer.save("C:/stress_model");

// evaluate the model using the user-independent "leave-one-user-our" method
Evaluation eval;
eval.evalLOUO(&trainer, allFeatures);
eval.print();
```

F Public Speaking Augmentation

F.1 SSJ Pipelines

The SSJ Pipeline for the public speaking social augmentation system discussed in Chapter 8 is split into two parts. First a small pipeline runs on the HMD (e.g. Google Glass) to collect the audio data from the microphone and stream it to a smartphone over Bluetooth. Then, a pipeline on the smartphone receives the audio data from the HMD and motion data from a Myo armband, and computes the speech rate and energy level of the user's behaviour. The results of this analysis are streamed back to the HMD, where a second SSJ pipeline will use it to generate and deliver feedback to the user as specified in the strategy file (see Appendix F.2). Theoretically, both pipelines could be merged and executed on the HMD if a sufficiently powerful device is available. If the Google Glass is used, its modest hardware specifications require the data processing steps to be delegated to a smartphone.

Pipeline for HMD

```
Pipeline pipe = Pipeline.getInstance();

//get audio data from microphone
Microphone mic = new Microphone();
AudioChannel audio = new AudioChannel();
audio.options.sampleRate.set(16000);
audio.options.scale.set(false);
pipe.addSensor(mic, audio);

//send it to the smartphone
BluetoothWriter socket = new BluetoothWriter();
socket.options.connectionName.set("audio");
socket.options.connectionType.set(BluetoothConnection.Type.CLIENT);
socket.options.serverAddr.set(PHONE_MAC_ADDRESS);
pipe.addConsumer(socket, audio, 0.1, 0);

//listen for analysis results
BluetoothEventReader receiver = new BluetoothEventReader();
receiver.options.connectionName.set("analysis");
receiver.options.connectionType.set(BluetoothConnection.Type.CLIENT);
receiver.options.serverAddr.set(PHONE_MAC_ADDRESS);
EventChannel analysisResults = pipe.registerEventProvider(receiver);

//give the feedback manager access to the analysis results
FeedbackManager feedback = new FeedbackManager(this);
feedback.options.strategy.set("strategy.xml");
pipe.registerEventListener(feedback, analysisResults);
```

Pipeline for Smartphone

```
Pipeline pipe = Pipeline.getInstance();

// Sensors
//receive audio data from HMD
BluetoothReader receiver = new BluetoothReader();
receiver.options.connectionName.set("audio");
receiver.options.connectionType.set(BluetoothConnection.Type.SERVER);
BluetoothChannel audio_raw = new BluetoothChannel();
audio_raw.options.outputClass.set(new String[]{"Audio"});
```

```
audio_raw.options.bytes.set(2);
audio_raw.options.dim.set(1);
audio_raw.options.type.set(Cons.Type.SHORT);
audio_raw.options.sr.set(16000.);
pipe.addSensor(receiver, audio_raw);

//prepare audio data for processing
AudioConvert audio = new AudioConvert();
pipe.addTransformer(audio, audio_raw, 0.1, 0);

//receive motion data from Myo armband
Myo myo = new Myo();
DynAccelerationChannel acc = new DynAccelerationChannel();
pipe.addSensor(myo, acc);

// Audio processing
//compute pitch
Pitch pitch = new Pitch();
pitch.options.detector.set(Pitch.YIN);
pitch.options.computePitch.set(false);
pitch.options.computePitchedState.set(false);
pitch.options.computeVoicedProb.set(true);
pipe.addTransformer(pitch, audio, 0.1, 0);

//filter the pitch
Envelope pitchf = new Envelope();
pitchf.options.attackSlope.set(0.3f);
pitchf.options.releaseSlope.set(0.05f);
pipe.addTransformer(pitchf, pitch, 1.0, 0);

//compute intensity (loudness) from audio data
Intensity intensity = new Intensity();
intensity.options.subtractMeanPressure.set(false);
pipe.addTransformer(intensity, audio, 1.0, 0);

//determine when user is talking to minimize resource consumption
ThresholdEventSender vad = new ThresholdEventSender();
vad.options.sender.set("SSJ");
vad.options.event.set("VoiceActivity");
vad.options.thresin.set(new float[]{40.0f}); //in dB
vad.options.mindur.set(1.0);
vad.options.maxdur.set(9.0);
vad.options.hangin.set((int) (intensity.getOutputStream().sr * 0.2)); //0.2s
vad.options.hangout.set((int) (intensity.getOutputStream().sr * 0.5)); //0.5s
Provider[] vad_in = {intensity};
pipe.addConsumer(vad, vad_in, 1.0, 0);
EventChannel vad_channel = vad.getEventChannelOut();

//compute "speech rate" from voiced segments
SpeechRate sr = new SpeechRate();
sr.options.sender.set("SSJ");
sr.options.event.set("SpeechRate");
sr.options.thresholdVoicedProb.set(0.3f);
sr.options.width.set(3);
Provider[] sr_in = {intensity, pitchf};
pipe.addConsumer(sr, sr_in, vad_channel);
```

```
EventChannel sr_channel = pipe.registerEventProvider(sr);

// Motion data processing
//compute the "energy" index
OverallActivation activity = new OverallActivation();
pipe.addTransformer(activity, acc, 0.1, 5.0);

//filter data
MvgAvgVar activityf = new MvgAvgVar();
activityf.options.window.set(10.);
pipe.addTransformer(activityf, activity, 0.1, 0);

//convert continuous stream to events
FloatsEventSender evactivity = new FloatsEventSender();
evactivity.options.sender.set("SSJ");
evactivity.options.event.set("OverallActivation");
pipe.addConsumer(evactivity, activityf, 0.5, 0);
EventChannel activity_channel = pipe.registerEventProvider(evactivity);

// Data output
//send analysis results to HMD for feedback generation
BluetoothEventWriter sender = new BluetoothEventWriter();
sender.options.connectionName.set("analysis");
sender.options.connectionType.set(BluetoothConnection.Type.SERVER);
pipe.registerEventListener(sender, sr_channel);
pipe.registerEventListener(sender, activity_channel);
```

F.2 Feedback Strategy

Using an SSJ feedback strategy, a multi-loop setup consisting of two behavioural feedback loops is defined. The first loop will generate visual feedback in response to the user's speech rate as computed by SSJ. The second loop will also deliver visual feedback, but in response to the user's body energy. In both cases, three feedback events are defined: low, medium and high intensity.

```
<ssj xmlns="hcm.ssj"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="hcm.ssj http://hcmlab.github.io/ssj/res/feedback.xsd">

 <strategy>

  <feedback type="visual" layout="layout_table" position="0">
   <condition type="SpeechRate" event="SpeechRate" sender="SSJ"
              history="5" from="0.0" to="2.0"/>
   <action res="orientation_low.png, thumb_negative.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table" position="0">
   <condition type="SpeechRate" event="SpeechRate" sender="SSJ"
              history="5" from="2.0" to="3.6"/>
   <action res="orientation_med.png, thumb_positive.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table" position="0">
   <condition type="SpeechRate" event="SpeechRate" sender="SSJ"
              history="5" from="3.6" to="999"/>
   <action res="orientation_high.png, thumb_negative.png"/>
  </feedback>

  <feedback type="visual" layout="layout_table" position="1">
   <condition type="BodyEnergy" event="OverallActivation" sender="SSJ"
              from="0.0" to="0.8"/>
   <action res="area_low.png, thumb_negative.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table" position="1">
   <condition type="BodyEnergy" event="OverallActivation" sender="SSJ"
              from="0.8" to="2.0"/>
   <action res="area_med.png, thumb_positive.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table" position="1">
   <condition type="BodyEnergy" event="OverallActivation" sender="SSJ"
              from="2.0" to="999"/>
   <action res="area_high.png, thumb_negative.png"/>
  </feedback>

 </strategy>
</ssj>
```

G Group Discussion Augmentation

G.1 SSJ Pipeline

To realize the group augmentation system described in Chapter 8, each user requires an Android device to run the behavioural feedback loop. For the visual feedback methods, the loop can be executed directly on the HMD or the tablet. In the case of the tactile and audio feedback methods, SSJ needs to run on a smartphone paired with a Myo armband (for tactile feedback) or a pair of headphones (for audio feedback). In all four cases, the same SSJ pipeline can be used, only the feedback strategies differ.

```
Pipeline pipe = Pipeline.getInstance();

// Sensors
//get audio data from microphone
Microphone mic = new Microphone();
AudioChannel audio = new AudioChannel();
audio.options.sampleRate.set(8000);
audio.options.scale.set(true);
pipe.addSensor(mic, audio);

// Audio processing
//compute intensity (loudness) from audio data
Intensity intensity = new Intensity();
intensity.options.subtractMeanPressure.set(false);
pipe.addTransformer(intensity, audio, 1.0, 0);

//determine when user is talking
ThresholdEventSender vad = new ThresholdEventSender();
vad.options.sender.set("SSJ");
vad.options.event.set("VoiceActivity");
vad.options.thresin.set(new float[]{40.0f}); //in dB
vad.options.mindur.set(1.0);
vad.options.maxdur.set(10.0);
vad.options.hangin.set((int) (intensity.getOutputStream().sr * 0.2)); //0.2s
vad.options.hangout.set((int) (intensity.getOutputStream().sr * 0.5)); //0.5s
pipe.addConsumer(vad, intensity, 1.0, 0);
EventChannel vad_channel = vad.getEventChannelOut();

// Feedback
FeedbackManager feedback = new FeedbackManager(this);
feedback.options.strategy.set("strategy.xml");
pipe.registerEventListener(feedback, vad_channel);
```

G.2 Feedback Strategies

The following code blocks show the four feedback strategies required for visual, auditory and tactile feedback delivery. Both visual devices (HMD and tablet) can use the same strategy.

Visual Feedback

```
<ssj xmlns="hcm.ssj"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="hcm.ssj http://hcmlab.github.io/ssj/res/feedback.xsd">

 <strategy>
  <feedback type="visual" layout="layout_table">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="0" to="30"/>
   <action res="100percent.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="30" to="60"/>
   <action res="75percent.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="60" to="90"/>
   <action res="50percent.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="90" to="108"/>
   <action res="25percent.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="108" to="120"/>
   <action res="10percent.png"/>
  </feedback>
  <feedback type="visual" layout="layout_table">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="120" to="999"/>
   <action res="timeover.png"/>
  </feedback>
</strategy>
</ssj>
```

Auditory Feedback

```
<ssj xmlns="hcm.ssj"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="hcm.ssj http://hcmlab.github.io/ssj/res/feedback.xsd">

 <strategy>
  <!-- events are only executed once -->
  <feedback type="audio">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="30" to="60"/>
   <action res="75percent.mp3" lockSelf="-1"/>
  </feedback>
```

```
  <feedback type="audio">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="60" to="90"/>
   <action res="50percent.mp3" lockSelf="-1"/>
  </feedback>
  <feedback type="audio">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="90" to="108"/>
   <action res="25percent.mp3" lockSelf="-1"/>
  </feedback>
  <feedback type="audio">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="108" to="120"/>
   <action res="10percent.mp3" lockSelf="-1"/>
  </feedback>
  <!-- event is executed once every second while the user speaks -->
  <feedback type="audio">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="120" to="999"/>
   <action res="timeover.mp3" lockSelf="1000"/>
  </feedback>
 </strategy>
</ssj>
```

Tactile Feedback

```
<ssj xmlns="hcm.ssj"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="hcm.ssj http://hcmlab.github.io/ssj/res/feedback.xsd">

 <strategy>
  <!-- events are only executed once -->
  <feedback type="tactile">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="30" to="60"/>
   <action intensity="150" duration="500" lockSelf="-1"/>
  </feedback>
  <feedback type="tactile">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="60" to="90"/>
   <action intensity="150,150" duration="500,500" lockSelf="-1"/>
  </feedback>
  <feedback type="tactile">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="90" to="108"/>
   <action intensity="150,150,150" duration="500,500,500" lockSelf="-1"/>
  </feedback>
  <feedback type="tactile">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="108" to="120"/>
   <action intensity="150" duration="500" lockSelf="-1"/>
  </feedback>
  <!-- event is executed once every second while the user speaks -->
  <feedback type="tactile">
   <condition type="SpeechDuration" event="VoiceActivity" sender="SSJ"
              sum="true" from="120" to="999"/>
   <action intensity="150" duration="2000" lockSelf="1000"/>
```

```
   </feedback>
  </strategy>
 </ssj>
```

H List of Personal Publications

Reference	Publication type	Personal role in project	Contribution to thesis
Damian et al. [2009]	Conference	First author, software development (character behaviour simulation)	informed general research direction
Endrass et al. [2010]	Conference	Co-author, software development (character behaviour simulation)	informed general research direction
Damian et al. [2011a]	Conference	First author, software development (software architecture, character behaviour simulation)	informed general research direction
Damian et al. [2011b]	Conference	First author, software development (character behaviour simulation), user study	informed general research direction
Damian [2012]	Diploma thesis	Sole author, software development (software architecture, character behaviour simulation), user study	informed Section 2.1
Obaid et al. [2012]	Conference	Co-author, user study, software development (augmented reality, character behaviour simulation)	informed general concept design
Kistler et al. [2012]	Journal	Co-author, software development (character behaviour simulation)	informed signal processing approach
Wagner et al. [2013]	Conference	Co-author, software development (social signal processing)	informed Chapter 7
Damian et al. [2013c]	Conference	First author, user study, software development (augmented reality, motion capturing, character behaviour simulation)	informed general research direction
Damian et al. [2013a]*	Workshop	First author, user study, software development (social signal processing)	informed signal processing approach
Porayska-Pomsta et al. [2013]**	Conference	Co-author, software development (social signal processing)	informed signal processing approach
Baur et al. [2013a]	Conference	Co-author, software development (social signal processing)	informed signal processing approach
Damian et al. [2013b]	Conference	First author, user study, software development (augmented reality, motion capturing, character behaviour simulation)	informed general research direction
Baur et al. [2013b]	Workshop	Co-author, software development (social signal processing)	informed Section 3.1
Anderson et al. [2013]	Conference	Co-author, software development (social signal processing)	informed Section 3.1
Porayska-Pomsta et al. [2014]	Conference	Co-author, software development (social signal processing), user study	informed signal processing approach
Jones et al. [2014]	Workshop	Co-author, software development (social signal processing)	informed signal processing approach
Gebhard et al. [2014]	Conference	Co-author, software development (social signal processing), user study	informed Section 3.1
Damian et al. [2014b]	Conference	First author, software development (social signal processing, live feedback generation), interaction design	informed general concept design
Damian et al. [2014a]	Workshop	First author, software development (social signal processing, live feedback generation), interaction design	informed general concept design
Damian et al. [2015b]	Conference	First author, software development (software architecture, social signal processing, live feedback generation), interaction design, user study	informed Chapters 7 and 8

Awadeen [2015]	Bachelor's thesis	Supervisor	informed Section 6.1.1
Damian et al. [2015a]	Conference	First author, user study, software development (social signal processing, game logic), interaction design	informed Section 3.1
Baur et al. [2015]	Journal	Co-author, software development (social signal processing), user study	informed Section 3.1
Damian and André [2016]	Conference	First author, software development (live feedback generation), user study	informed Section 6.1.3
Dietz et al. [2016]	Conference	Co-author, project supervision, user study	informed general concept design
Damian et al. [2016a]	Conference	First author, software development (software architecture, social signal processing, multimodal feedback), user study	informed Chapters 7 and 9
Damian et al. [2016b]	Conference	First author, software development (software architecture, social signal processing, multimodal feedback)	informed Chapter 7
Bottari [2017]	Master's thesis	Supervisor	informed Section 6.2.2

* Won best paper award
** Won most participative demo award